UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

ENCODAGE DES SIGNAUX DE PAROLE PAR INVERSION DES MOTIFS D'EXCITATION AUDITIVE

Thèse de doctorat
Spécialité : génie électrique

Khaled LAKHDHAR

Sherbrooke (Québec) Canada

Décembre 2017

MEMBRES DU JURY

Roch LEFEBVRE

Directeur

Éric PLOURDE

Évaluateur

Jean ROUAT

Évaluateur

Martin BOUCHARD

Évaluateur

RÉSUMÉ

Le système auditif humain est, comme tout autre système sensoriel biologique, complexe et redondant. Cette thèse avance la revendication selon laquelle il est possible, à faible complexité, de compresser et de synthétiser un signal sonore exclusivement dans le domaine perceptuel. On traite l'approche du codage par inversion des représentations bio-inspirées et on propose un codeur qui exploite la redondance présente dans ces représentations. Le codeur proposé transforme ces derniers en une représentation compacte dépourvue de redondance perceptuelle ce qui facilite la compression tout en permettant une bonne qualité subjective de reconstruction.

La première partie de cette thèse propose un nouveau filtre auditif à faible complexité qui peut non seulement modéliser les réponses mécaniques de la membrane basilaire, synthétiser les réponses impulsionnelles du nerf auditif mais aussi expliquer les expériences du masquage fréquentiel. La deuxième partie détaille l'exploitation de ce même banc de filtres auditifs pour la discipline de la compression des signaux sonores. Des modèles de masquage adaptés à ce banc de filtres sont appliqués aux motifs d'excitation auditives pour obtenir des représentations éparses. Des expériences montrent que ce codeur permet de réduire considérablement la redondance dans le domaine perceptuel tout en maintenant une bonne qualité subjective de synthèse.

Mots-clés : filtres auditifs, distribution binomiale, masquage, parcimonie, psychoacoustique, compression, codage audio

REMERCIEMENTS

Je tiens à remercier en premier lieu mon directeur de recherche Pr. Roch Lefebvre pour m'avoir accueilli parmi le groupe de recherche GRPA (groupe de recherche sur la parole et l'audio). Je lui suis aussi reconnaissant pour ses encouragements et judicieux conseils. Je remercie aussi Pr. Jean Rouat et Pr. Roger Goulet pour avoir fait partie de mon comité de conseil. Leurs questions durant l'examen de synthèse m'ont permis de mieux cerner la question de recherche à laquelle cette thèse répond. Mes remerciements s'adressent aussi à Danielle Poirier, secrétaire du groupe, pour m'avoir maintes fois aidé avec les procédures administratives, pour sa bonne humeur et sa sympathie.
Je tiens aussi à remercier plusieurs personnes de la communauté scientifique notamment Richard F. Lyon pour avoir partagé les données expérimentales collectées par Glasberg et Moore et Roland Carrat pour avoir partagé son ouvrage « l'oreille numérique » avant sa publication.

Je ne peux finir sans remercier aussi la Mission Universitaire de Tunisie en Amérique du Nord pour sa présence tout au long de mon doctorat et pour son support aussi bien financier que moral.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DE VARIABLES

$\mathbf{b_0}$ Vecteur contenant les ordonnées à l'origine utilisées pour modéliser les réponses impulsionnelles de la membrane basilaire de chats.

β Paramètre contrôlant la position des pôles de la fonction de transfert donnée par l'équation (2.15).

$e(\mathbf{f})$ Erreur quadratique moyenne de modélisation des réponses impulsionnelles de la membrane basilaire de chats résultant de l'utilisation d'un modèle $\mathbf{f}$.

f_0 fréquence centrale de la tonalité utilisée pour les expériences du masquage fréquentiel.

$\mathbf{f}$ Relation linéaire utilisée pour modéliser les réponses impulsionnelles de la membrane basilaire de chats.

G_i i-ième filtre de synthèse du banc de filtres.

γ Paramètre contrôlant la position des zéros de la fonction de transfert donnée par l'équation (2.15).

H_i i-ième filtre d'analyse du banc de filtres.

$\mathcal{H}$ transformée de Hilbert.

∇ le Jacobien d'une fonction differentiable.

K constante de détection dans l'équation (3.2).

$\mathcal{L}$ transformée de Laplace.

$\mathbf{M}$ Modèle utilisé pour modéliser le filtre auditif.

$\mathbf{M}$ Vecteur contenant les coefficients directeurs utilisés pour modéliser les réponses impulsionnelles de la membrane basilaire de chats en fonction de P_N.

n Ordre du filtre binomial tel que décrit par l'équation (2.15).

$\tau_{\mathbf{éch}}$ Nombre d'impulsions par échantillons audio (équation (5.20)).

N_0 niveau du bruit utilisé pour les expériences du masquage fréquentiel.

P_0 niveau du bruit interne de fond au niveau de l'oreille.

P_N Niveau d'excitation sonore utilisé pour recueillir les réponses impulsionnelles de [Carney *et coll.*, 1999].

P_s niveau de la tonalité utilisée pour les expériences du masquage fréquentiel.

$\hat{P}_s$ niveau de la tonalité prédit par le modèle dans le cas des expériences du masquage fréquentiel.

P_x Niveau d'excitation sonore normalisé donné par P_N-80.

$\hat{RI}_{\mathbf{f}}(i)$ I ème réponse impulsionnelle générée utilisant un modèle $\mathbf{f}$ conformément à l'équation (2.18).

$s(n)$ signal audio à l'entrée du banc de filtres d'analyse.

$\hat{s}(n)$ signal audio à la sortie du banc de filtres de synthèse.

τ_s Taux de parcimonie après classification des impulsions en impulsions masquantes et masquées (équation (5.19)).

$\tau_{\mathbf{cmp}}$ taux de compression moyen des réponses impulsionnelles donné par l'équation (2.20).

w_0 fréquence de résonance du filtre auditif.

$\xi(M)$ erreur de prédiction des tonalités dans le cadre des expériences de masquage fréquentiel en utilisant un modèle M.

$y_i(n)$ signal audio à l'entrée du i-ième filtre du banc de filtres.

$\hat{y}_i(n)$ signal audio à la sortie du i-ième filtre du banc de filtres.

$\mathcal{Z}$ transformée en z.

LISTE DES ACRONYMES

AGC filtre gammachirp analytique, *analytical gammachirp filter*.

APFC tout-pôle en cascade, *all-pole filter cascade*.

APG gammatone tout-pôle, *all-pole gammatone filter*.

BIT filtre binomial, *Binomial-tone filter*.

BIT$_2$ filtre binomial d'ordre 2.

BIT$_3$ filtre binomial d'ordre 3.

BIT$_n$ filtre binomial d'ordre n, *Binonial-tone filter*.

BWT transformation de Burrows-Wheeler.

cBIT filtre binomial compressif, *Compressive Binonial-tone filter*.

cBIT$_2^*$ filtre binomial compressif d'ordre 2 réalisant le compromis entre erreur d'apprentissage et de généralisation.

cBIT$_2$ filtre binomial compressif d'ordre 2.

cBIT$_3$ filtre binomial compressif d'ordre 3.

cBIT$_n$ filtre binomial compressif d'ordre n.

cGC filtre gammachirp compressif, *compressive gammachirp filter*.

CI corrélation inversée.

CS acquisition comprimée, *compressive sampling*.

DAPG gammatone tout-pôle différencié, *differenciated all-pole gammatone*.

DPCM codage par modulation des différences, *Differential pulse code modulation*.

DRNL *Dual Resonance Nonlinear Filterbank*.

EQM l'erreur quadratique moyenne.

ERB largeur de bande rectangulaire équivalente, *equivalent rectangular bandwidth* .

FI fréquence instantanée.

GC filtre gammachirp, *gammachirp filter*.

GT filtre gammatone, *gammatone filter*.

MA moyenne mobile, *moving average*.

MB membrane basilaire.

MDCT transformée en cosinus discrète modifiée, *modified discrete cosine transform*.

MF meilleure fréquence, *best frequency*.

MP *matching pursuit*.

NA nerf auditif.

ODG différence de qualité objective, *objective difference grade*.

OZFC tout-pôle en cascade avec un zéro, *one-zero filter cascade*.

OZG gammatone à un zéro, *one-zero gammatone filter*.

PEAQ évaluation perceptuelle de la qualité audio, *Perceptual evaluation of audio quality*.

PZFC pôles-zéros en cascade, *pole-zero filter cascade*.

PZFC5 pôles-zéros en cascade où les pôles et zéros se déplacent par le même taux.

RI réponse impulsionnelle.

RIF filtre à réponse impulsionnelle finie.

RII filtres à réponse impulsionnelle infinie.

RII réponse impulsionnelle infinie.

RLE codage par plage (*run length coding*).

RSB rapport signal sur bruit.

SDG différence de qualité subjective, *subjective difference grade*.

TFD transformée de Fourier discrète.

TIMIT collection de signaux de parole collectés par Texas Instruments et Massachusetts Institute of Technology.

WFT transformée de Fourier enveloppée, *wrapped Fourier transform*.

CHAPITRE 1

Introduction

1.1 Mise en contexte

L'introduction du disque compact (CD, *Compact Disc*) au début des années quatre-vingt, a permis aux utilisateurs d'apprécier l'avantage de la représentation numérique du signal sonore étant donné que cette représentation est robuste et permet de conserver une bonne qualité de reconstruction. Ces avantages, cependant, ont été obtenus au détriment d'un débit très élevé. Les disques compacts conventionnels et les rubans audionumériques (DAT, *Digital Audio Tape*) sont typiquement échantillonnés à une fréquence de 44.1 kHz ou 48 kHz avec une résolution de 16 bits. Ces configurations, pour des signaux monophoniques, donnent des débits très élevés d'encodage valant 705.6 kbps (kilo bits par seconde) par canal pour une fréquence d'échantillonnage de 44.1 kHz et 768 kbps par canal pour une fréquence d'échantillonnage de 48 kHz [Painter et Spanias, 2000].

Même très élevés, ces débits d'encodage ont été bien adaptés aux applications multimédias première-génération tels que le DAT ou le CD. Malheureusement, les applications multimédias seconde-génération, et les systèmes sans fil particulièrement, sont souvent sujets à des restrictions relatives aux bandes passantes ou bien aux coûts de stockage. Ces contraintes font en sorte que des algorithmes de compression sont inévitables. Suite au succès connu par les applications multimédia première-génération (CD et DAT ...), les utilisateurs s'attendent à une « qualité-CD » pour tout média reproduis. Pour cette raison, les nouveaux systèmes d'encodage audio doivent réduire les débits d'encodage sans compromettre la qualité de reproduction.

Ces considérations ont motivé une recherche intensive dont le but est de formuler et réaliser des schémas d'encodage qui peuvent satisfaire des demandes conflictuelles d'un bas débit et d'une reproduction transparente. La transparence à l'écoute implique que les utilisateurs ne pourraient discerner le signal original du signal encodé. De ce fait, les techniques recherchées sont celles qui peuvent assurer une transparence à l'écoute et non forcément une meilleure conservation du rapport signal au bruit par exemple. L'oreille humaine présente des limites que ce soit par rapport à sa résolution fréquentielle ou temporelle [Andoh *et coll.*, 2005; Békésy, 1953; Hartmann *et coll.*, 2010; Koike *et coll.*, 2005; Kollmeier *et coll.*, 2008; Matsui *et coll.*, 2006; Pratt et Sohmer, 1976; Wada *et coll.*, 2002]. Des phénomènes

de masquage fréquentiel et temporel font en sorte que certains sons ne sont pas perçus. Cette inaptitude de l'oreille à discerner des sons et plus particulièrement les bruits en présence d'autres sons masquants est la clé du développement des codeurs avec perte mais psychoacoustiquement transparents. Avec perte, parce que le signal encodé a un rapport signal sur bruit non infini, transparents parce qu'entre le signal original et celui encodé, la différence à l'écoute est souvent difficilement audible. Le domaine de la psychoacoustique a fait de grands pas envers la caractérisation de l'oreille humaine. Dans le but d'étudier le fonctionnement de l'oreille plusieurs expériences ont été développées [Moore, 1987; Zwicker *et coll.*, 1982; Zwicker et Terhardt, 1974]. Même si les résultats de ces études sont disponibles, les synthétiser pour en faire un modèle générique de l'oreille interne se révèle être une tâche difficile. En effet, différents phénomènes doivent être pris en considération : la haute non-linéarité de l'oreille interne, l'étalement de la réponse de la membrane basilaire, les notions de battements... De ce fait, les algorithmes d'encodage sont contraints de compter sur des modèles simples, souvent simplificateurs et imprécis [Brandenburg, 1999; Morris, 1995; Painter et Spanias, 2000]. Les modèles actuellement adoptés sont souvent sujets à certaines hypothèses simplificatrices telles que l'additivité des masquants, la linéarité de l'oreille interne quant aux excitations à faibles niveaux et la nature des signaux masquants limitée aux bruits blanc et aux tonalités. Bien que l'application de règles perceptuelles à l'encodage des signaux audio n'est pas une nouvelle idée, la plupart des codeurs récents réalisent la compression en exploitant le fait que l'information « inutile » est indétectable par l'oreille. L'information inutile est généralement identifiée durant une étape d'analyse, et ce, en incorporant dans le corps de l'encodeur plusieurs principes psychoacoustiques tels que le seuil d'audition absolue, l'analyse par bandes critiques et le masquage simultané. La combinaison de ces principes, avec ceux de la quantification, a mené aussi au développement de l'entropie perceptuelle [Painter et Spanias, 2000], une estimation quantitative de la limite théorique de la compression transparente des signaux audio.

Plusieurs familles de codeurs audio ont fini par intégrer un module psychoacoustique dont la fonction est de contrôler l'allocation du débit en fonction de la pertinence perceptuelle de l'information à transmettre. Une schéma bloc d'une telle approche est donné à la figure 1.1.

Dans ces codeurs, le signal audio est projeté dans deux espaces différents. Le premier est souvent un espace compacte où le signal est transformé par example dans le domaine de la transformée de Fourier ou la transformée en cosinus discrète modifiée, *modified discrete cosine transform* (MDCT) [Britanak et Rao, 2001]. Ceci a pour effet de représenter le signal

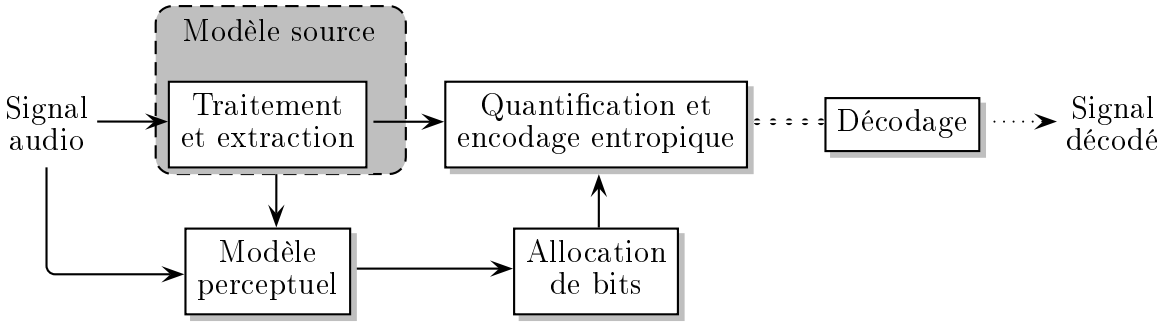


Figure 1.1 Schéma bloc d'un codeur audio.

en un ensemble d'éléments faciles à compresser. Le deuxième espace est une approximation du domaine perceptuel par example en utilisant la transformée de Fourier enveloppée, *wrapped Fourier transform* (WFT) où par filtrage sous-bande étalé sur une échelle mimant la sélectivité fréquentielle de l'oreille humaine [Feldbauer et Kubin, 2003; Makur et Mitra, 2001]. Le bloc de l'allocations perceptuelle de bit implémente des règles perceptuelles et dicte le mode d'opération du bloc de quantification.

Même si un éventail de techniques est présent et détaillé dans la littérature, rares sont celles qui ont pu devenir des standards internationaux ou commerciaux car souvent d'autres contraintes s'additionnent à l'exigence de la transparence. Ces contraintes sont d'autant plus importantes et limitent encore les codeurs quant aux choix des techniques d'analyse ou de traitement. Les plus importantes sont la vitesse d'encodage et de décodage, le délai algorithmique et l'occupation de la mémoire.

La figure 1.2 présente une coupe transversale de l'oreille chez les humains. Le tout com-

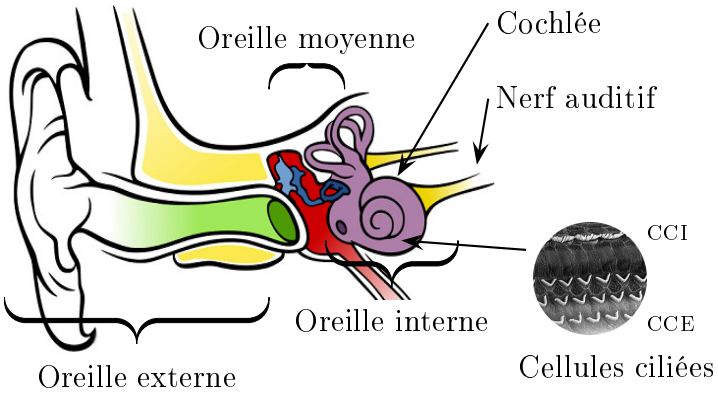


Figure 1.2 Coupe transversale de l'oreille [Teachmeanatomy, 2016].

mence quand l'onde acoustique atteint l'oreille externe. Cette onde traverse le conduit

auditif et vient frapper contre le tympan. Ces vibrations sont transmises par les osselets au liquide se trouvant à l'intérieur de la cochlée. À l'intérieur de la cochlée, la membrane basilaire réagit par des vibrations localisées. Ce mouvement est détecté par les cellules ciliées internes qui le transforment en fluctuations de potentiel électrique contrôlant ainsi la libération des neurotransmetteurs au niveau de la connection synaptique. De ce fait des potentiels d'actions sont générés au niveau de plusieurs fibres du nerf auditif qui transmet cette information au tronc cérébral ensuite au cortex auditif. Une schématisation possible de ce fonctionnement de l'oreille peut être représentée par la figure 1.3 où l'action du système nerveux efférent est représentée par des lignes discontinues.

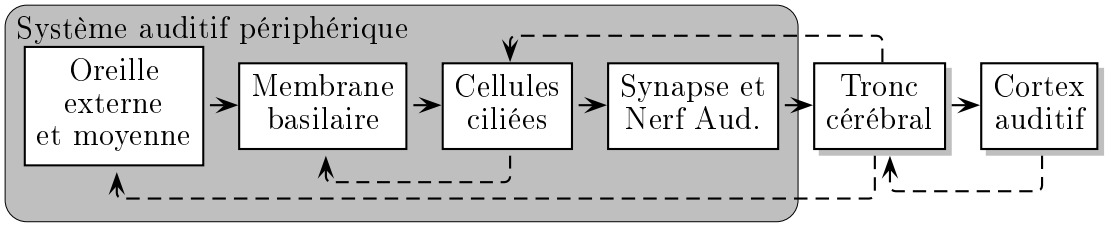


Figure 1.3 Modélisation schématique du système auditif biologique.

Le tableau 1.1 décrit les modèles computationnels souvent utilisés pour modéliser les différents bloc de la figure 1.3.

Oreille ext/moy	**Cochlée(MB)**	**Cellules cil.**	**CCI-FNA**
•Filtre IIR	•GammaTone	•Non-linéarités	•Trois réservoirs
•Filtre FIR	•GammaChirp	•Filtres pass-bas	•Équ. diff
	•PZFC	•Intégrateur à fuite	
	•CAR-FAC		
	•Carney et al.		
	•DRNL		

Tableau 1.1 Modèles computationnels du système auditif périphérique.

L'action de l'oreille moyenne et externe est souvent modélisée par un filtre linéaire. L'action de la membrane basilaire est modélisée quant à elle par banc de filtres en parallèle pour simuler la tonotopie de la cochlée. Parmi les bancs de filtres populaires on peut citer par examples les filtres Gammatones [Patterson, 1986], les filtres Gammachirps [Irino et Patterson, 2006a] ou les filtres *Dual Resonance Nonlinear Filterbank*s (DRNLs) [Meddis et O'Mard, 2005].

Les cellules ciliées internes sont responsables de la transconductance méchano-électrique dans l'organe de Corti. Leurs actions sont estimées expérimentalement comme étant le rapport entre la composante continue et la composante alternative de leurs réponses aux

différents stimuli. Elles sont souvent modélisées comme étant une cascade de gains de saturation suivie par un filtre pass-bas [Karjalainen, 1987; Meddis et Lopez-Poveda, 2010; Meddis et O'Mard, 2005].

La libération de neurotransmetteurs au niveau de la connection entre les cellules ciliées et les fibres du nerf auditif est un processus stochastic. La probabilité instantanée décrivant cette libération est function de la concentration du Calcium et le nombre des vésicules disponibles. Meddis et O'Mard [2005] modélisent cette connection par des équations différentielles mais comme modéliser les impulsions individuelles est souvent computationnellement très coûteux, Zilany *et coll.* [2009] proposent un modèle où cette connection est modélisée comme une somme de lois de puissance à deux constantes de temps et un processus gaussien fractionnaire alimentant un processus de Poisson.

Dans le but d'obtenir un encodeur bio-inspiré perceptuellement transparent, un modèle numérique fidèle du système auditif périphérique humain s'impose. Parvenir à modéliser numériquement une vraie oreille humaine avec des coûts computationnels raisonnables est une tâche difficile. Même si cela est possible, exploiter ce modèle pour compresser les signaux audio reste encore une question sans réponses concluantes. On propose de simplifier le schéma de la figure 1.3.

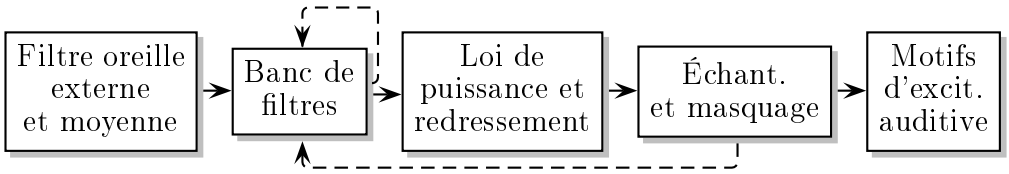


Figure 1.4 Modèle auditif proposé.

Dans le modèle de la figure 1.4, l'action de la membrane basilaire est modélisée par un banc de filtres en peigne alors que les cellules ciliées internes sont modélisées par une loi de puissance suivie d'un redressement simple alternance. L'action des synapses et du nerf auditif est modélisée par un échantillonnage adaptatif. Alors que cette architecture n'est pas nouvelle en tant que telle des problèmes reliés à la complexité d'implémentation et aux délais des traitements restent encore non résolus limitant ainsi l'exploitation de cette approche pour la discipline du codage audio.

1.2 Questions de recherches et contributions originales

Dans cette thèse on étudie la possibilité de réaliser un codeur audio opérant dans le domaine perceptuel par extraction et compression des motifs d'excitation neuronale du

système auditif. Contrairement aux codeurs basés sur un modèle source, les opérations de codage se font dans le domaine perceptuel : le signal est transformé en motifs d'excitation auditive épars. Une fois dans ce domaine perceptuel, des modèles de masquage simultané et temporel sont utilisés pour éliminer la redondance perceptuelle. Ces motifs d'excitation réduits sont ensuite compressés pour former un flux binaire. Le décodeur consomme ce flux binaire et reconstruit le signal original par inversion des motifs d'excitation auditive. Cette approche est différente de celles souvent adoptées par les codeurs cités plus haut où un modèle psychoacoustique est utilisé seulement pour dicter l'allocation du débit. Un schéma bloc d'une telle approche est donné par la figure 1.5.

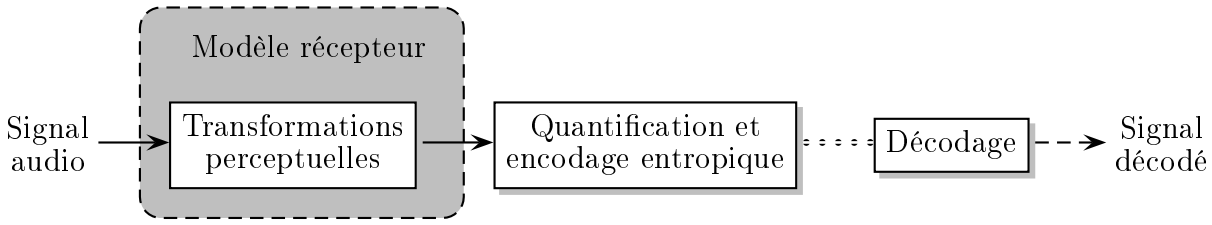


Figure 1.5 Codage proposé dans le domaine perceptuel.

Souvent les approches visant la compression dans le domaine perceptuel se heurtent à plusieurs défis [Feldbauer, 2005; Thiemann, 2011]. Le premier défi concerne les transformations nécessaires à la représentation d'un signal dans le domaine perceptuel : Quelles transformations simples en implémentation peuvent être utilisées pour produire des motifs d'excitation auditive ? Le deuxième défi touche à la synthèse du signal audio à partir de ses représentations perceptuelles : Quelles approches simples en implémentation suivre pour synthétiser le signal audio à partir de ces motifs ? Le troisième défi concerne la compression entropique de ces motifs où on se pose la question suivante : Jusqu'à quel taux de compression des motifs d'excitation auditives peut on espérer tout en maintenant une bonne qualité de synthèse ? Cette thèse fournit des réponses à ces questions et montre qu'il est possible de réaliser du codage par inversion des motifs d'excitation auditive avec une complexité réduite à moyen débit sans dégradations perceptibles .Pour aboutir à ce résultat, cette thèse est organisée en trois grandes parties chacune répondant aux questions de recherche.

On commence dans la première partie par aborder la complexité souvent citée dans la littérature quand il s'agit de modéliser le système auditif humain. Le banc de filtres auditifs étant souvent la partie la plus complexe à implémenter, le chapitre 2 introduit l'idée originale des filtres binomiaux et prouve qu'ils peuvent être utilisés pour modéliser les réponses impulsionnelles du nerf auditif chez les chats collectées par [Carney *et coll.*,

1999]. Une comparaison détaillée avec l'état de l'art est aussi donnée dans le même chapitre pour montrer l'originalité de ces filtres et leurs complexités d'implémentation réduites. Le chapitre 3 montre que cette même famille de filtres peut aussi expliquer les phénomènes du masquage fréquentiel chez les humains. On montre en utilisant les expériences du masquage des tonalités par du bruit blanc à bandes étroites de [Baker *et coll.*, 1998; Glasberg et Moore, 2000] que les filtres binomiaux fournissent d'excellents modèles pour prédire les résultats de telles expériences. On compare aussi ces filtres aux familles de filtres auditifs populaires détaillés dans la littérature : encore une autre fois les filtres binomiaux se distinguent par leurs complexités d'implémentation réduites pour les mêmes performances quand il s'agit de prédire le seuil de masquage des tonalités par un bruit blanc à bandes étroites.

Le chapitre 4 introduit la notion de synthèse par inversion des motifs d'excitation auditives et propose une nouvelle approche d'égalisation permettant une reconstruction parfaite du signal. Cette approche est très simple à implémenter et permet d'éviter la complexité inhérente aux approches d'analyse par synthèse appliquées au problème d'égalisation ou celles basées sur des recherches exhaustives. Le chapitre 5 introduit des algorithmes efficaces de masquage dans le domaine perceptuel. L'application de ces algorithmes résulte en une réduction du nombre d'impulsions et permet en utilisant un seul paramètre de contrôle d'ajuster ce nombre à la qualité de synthèse souhaitée. Pour compenser la perte d'énergie due à la mise des impulsions masquées à zéro, un algorithme adaptatif de correction est proposé. Cet algorithme, opérant en boucle ouverte, permet de restaurer cette perte d'énergie avec une coût computationnel réduit.

Les motifs d'excitation auditives étant épars, dans le chapitre 6 on introduit des algorithmes de compression avec et sans perte de ces motifs. On opte pour une approche où les positions des impulsions masquantes sont compressées sans perte alors que leurs amplitudes sont quantifiées de façon grossière. Les résultats obtenus confirment qu'il est possible avec cette approche d'obtenir des taux de compression compétitifs tout en maintenant une bonne qualité subjective de synthèse. Finalement le chapitre 7 résume les travaux originaux présentés dans cette thèse et expose les défis de la compression des signaux épars en proposant des pistes permettant de contourner ces obstacles.

Cette thèse présente les algorithmes, les résultats de simulation et les conclusions obtenues permettant de concevoir et implémenter un codeur audio joignant la discipline de la modélisation auditive à celle de la compression audio. On parvient donc à répondre aux questions de recherche citées plus haut : Il est possible de réaliser du codage audio à large

bande par inversion des motifs d'excitation auditive et ce à moindre coût computationnel et à moyen débit tout en maintenant une bonne qualité subjective de synthèse.

CHAPITRE 2

Les filtres binomiaux : un modèle original des réponses impulsionnelles du nerf auditif à faible complexité

La réponse impulsionnelle (RI) de n'importe quel système linéaire peut être estimée en utilisant la corrélation croisée. Ceci peut être réalisé en utilisant la corrélation croisée entre la réponse du système et la forme d'onde du bruit à large bande qui a généré cette réponse. Quand le signal d'excitation est un bruit blanc (stationnaire et ergodique), la RI peut être estimée directement [De Boer et De Jongh, 1978]. Cette technique est utilisée comme une estimation indirecte de la composante linéaire des réponses de la membrane basilaire (MB) alors que la réponse à une excitation sous forme de clic est une estimation directe [de Boer et Nuttall, 1997]. La corrélation inversée (CI) est une extension de la méthode de corrélation croisée et est utilisée comme une estimation indirecte de la composante linéaire de la RI du nerf auditif (NA) [de Boer et de Jongh, 1978]. Plus de détails concernant la technique de la CI peuvent être trouvés dans [Dayan et Abbott, 2002; Theunissen *et coll.*, 2001].

Une fréquence instantanée (FI) variable est présente dans les RIs de la MB ainsi que celles des fibres du nerf auditif [Carney *et coll.*, 1999; de Boer et Nuttall, 1997; Tan et Carney, 2003]. Les RIs des fibres ont des FIs dont la trajectoire est indépendante du niveau d'excitation. C'est à dire que les temps de passage par zéro des RIs sont indépendants des niveaux d'excitation. Cette trajectoire est croissante pour des fréquences supérieures à 1.5 kHz, relativement constante pour des fréquences comprises entre 750 Hz et 1.5 kHz et a un taux de glissement décroissant pour les fréquences inférieures à 750 Hz. Ce glissement affecte non seulement la structure fine de la réponse du NA, mais également la déviation de la meilleure fréquence, *best frequency* (MF) en fonction du niveau de la pression acoustique.

Les FIs des RIs du NA peuvent être estimées en utilisant la transformée de Hilbert par example. Pour un signal réel $s(t)$, le signal analytique $s_a(t)$ est donné par [Boashash, 1992] :

$$s_a(t) = s(t) + i\hat{s}(t) = A(t)e^{i\psi(t)} \tag{2.1}$$

Où $\hat{s}(t)$ est la transformée de Hilbert du signal $s(t)$. On peut alors dans ce cas estimer la FI comme étant la dérivée de la phase du signal analytique :

$$fi(t) = \frac{d\psi(t)}{dt} \tag{2.2}$$

La trajectoire des FIs est donnée comme étant la pente qui décrit l'évolution de la fréquence instantanée fi au cours du temps.

Différents modèles ont essayé de simuler la RI [1] du NA tout en prenant en compte de l'enveloppe qui ressemble à une distribution gamma, de la trajectoire des FIs et de la compression observée autour de la meilleure fréquence.

Le modèle proposé par [Carney *et coll.*, 1999; Tan et Carney, 1999] se compose d'un filtre du 11 ème ordre simulant l'oreille moyenne mis en cascade avec un filtre contrôlé par une fonction non linéaire combinée à un mécanisme de rétroaction afin d'introduire une compression non-linéaire. Ce dernier filtre a deux pôles de huitième ordre et un pôle de quatrième ordre, leurs complexes conjugués et un zéro purement réel du 11 ème ordre. Cette architecture a pour but de simuler la variation de la forme du filtre auditif en fonction du niveau de simulation tout en maintenant des trajectoires stables des fréquences instantanées. Zilany et al. dans [Zilany et Bruce, 2006] réduisent de moitié le nombre des coefficients et étendent le modèle à des niveaux sonores élevés.

Irino et Patterson dans [Irino et Patterson, 1997] ont proposé un filtre auditif appelé filtre gammachirp, *gammachirp filter* (GC). Le filtre gammachirp analytique, *analytical gammachirp filter* (AGC) a été présenté comme une extension du filtre gammatone, *gammatone filter* (GT) et a été le premier modèle à modéliser explicitement la trajectoire des FIs. Dans [Irino et Patterson, 2001], l'architecture du GC a été revue afin de prendre en compte les résultats publiés par [Carney *et coll.*, 1999]. Le filtre gammachirp compressif, *compressive gammachirp filter* (cGC) est constitué d'un filtre GT en cascade avec une fonction asymétrique passe-bas cascadée à un autre filtre passe-haut dont le gain dépend du niveau d'excitation. L'implémentation numérique du GC compressif a été réalisée comme un filtre à réponse impulsionnelle infinie (RII) utilisant 80 coefficients par chaîne auditive [Irino et Patterson, 2006a].

Le modèle proposé par [Lyon, 2011a] a été inspiré par la méthode Wentzel-Kramers-Brillion utilisée pour trouver des solutions approximatives des équations aux dérivées partielles en

1. Les systèmes étudiés dans cette thèse sont non-linéaires. Cependant quand le niveau d'excitation est constant, les bancs de filtres de ces systèmes sont considérés linéaires et peuvent donc être caractérisés par leurs réponses impulsionnelles.

milieu semi-clos. Dans son modèle, la réponse du nerf auditif en une position donnée sur la cochlée est équivalente à la réponse d'une séquence de filtres mis en cascade. Chaque étage se compose d'un filtre biquadratique de second ordre où la position des zéros et des pôles est fonction du niveau d'excitation. Afin de préserver un taux de glissement indépendant du niveau d'excitation, les pôles et les zéros sont limités à des déplacement proportionnels. Avec cette dernière contrainte, seules les FIs avec les taux de glissement croissants peuvent être modélisés.

La plupart des modèles du NA ont adopté la distribution gamma comme étant un modèle représentatif de l'enveloppe de la RI du NA. Le filtre GT (l'un des moins complexes en terme de description et implémentation) par exemple, est défini comme étant une tonalité multipliée par une enveloppe ressemblant une distribution gamma. Même si la description temporelle de cette dernière est compacte, la representation fréquentielle du filtre gammatone est complexe ce qui limite la possibilité d'une implementation digitale efficace.

Dans ce chapitre, les filtres binomiaux sont introduits comme étant des alternatives moins complexes aux filtres GTs et GCs. De plus, par placement judicieux des zeros de la fonction de transfert des filtres binomiaux, un glissement des FIs peut être introduit dans la RI de ces derniers. Le modèle proposé produit des réponses réalistes qui sont en accord avec les données physiologiques : la réponse temporelle du modèle ressemble à la RI du NA, la trajectoire de la FI est indépendante du niveau d'excitation et la fonction entrée-sortie du modèle est compressive.

2.1 Les filtres binomiaux

2.1.1 Motivation

La RI du NA peut être décomposée comme étant une enveloppe multipliée par une tonalité. La forme d'une loi gamma a été souvent utilisée comme un modèle de l'enveloppe de la RI des fibres auditives. Dans ce cas :

$$M_{NA}(t) = t^{n-1} \times \exp(-\gamma t) \times M_{por}(t) \tag{2.3}$$

Où M_{NA} représente le modèle de la réponse impulsionnelle du NA et M_{por} représente le modèle de la tonalité. Dans le cas du filtre GT d'ordre n, la porteuse est une simple sinusoïde :

$$\mathrm{GTF}(t) = t^{n-1} \times \exp(-\gamma t) \times \cos(\omega t + \phi) \tag{2.4}$$

Même si la transformée en s d'un oscillateur amorti est simple, le terme t^{n-1} se transforme en une dérivée dans le domaine de la transformée de Laplace, ce qui complique la fonction de transfert du filtre gammatone.

La fonction de transfert du filtre GT d'ordre n (équation (2.4)) a été donnée par [Katsiamis *et coll.*, 2007] :

$$H_{\mathrm{GTF}}(s) = \frac{e^{j\phi}[s+s_0]^n + e^{-j\phi}[s+\overline{s_0}]^n}{[(s+b)^2+w_0^2]^n} \tag{2.5}$$

Où le zéro $s_0 = w_0/2Q + jw_0\sqrt{1-1/4Q^2}$, w_0 la fréquence centrale du filtre auditif et Q son facteur de qualité à $-3dB$. À cause de cette description très complexe, Lyon a présenté dans [Lyon, 1996] un filtre appelé gammatone tout-pôle, *all-pole gammatone filter* (APG) où tous les zéros ont été retirés de l'équation (2.5).

Dans [Katsiamis *et coll.*, 2007], les auteurs ont présenté un modèle basé sur le filtre APG : ils ont introduit un zéro dans la fonction de transfert de ce dernier ce qui donna deux familles de filtres : le gammatone tout-pôle différencié, *differenciated all-pole gammatone* (DAPG) et gammatone à un zéro, *one-zero gammatone filter* (OZG). En terme de paramétrisation cartésienne, la transformée de Laplace de ces deux filtres est donnée par :

$$H_{\mathrm{DAPGF}}(s) = \frac{Ks}{[(s+b)^2+w_0^2]^N} \tag{2.6}$$

$$H_{\mathrm{OZGF}}(s) = \frac{K(s+w_z)}{[(s+b)^2+w_0^2]^N} \tag{2.7}$$

Où N est l'ordre du filtre, K une constante et w_z le zéro additionnel. Les auteurs ont aussi considéré deux versions en cascades de ces filtres : le filtre tout-pôle en cascade, *all-pole filter cascade* (APFC) et le filtre pôles-zéros en cascade, *pole-zero filter cascade* (PZFC) dont la fonction de transfert est donnée par.

$$H_{\mathrm{PZFC}}(s) = \prod_{k=0}^{N} \frac{[(s+z_k)^2+w_0^2]}{[(s+p_k)^2+w_0^2]} \tag{2.8}$$

Où N est le nombre de chaînes par largeur de bande rectangulaire équivalente, *equivalent rectangular bandwidth* (ERB) souvent fixé à 2 ou 3 par ERB [Katsiamis *et coll.*, 2007].

Dans le but d'éviter la dérivée introduite par le terme t^{n-1}, on présente dans les sections suivantes le filtre binomial comme étant une alternative moins complexe que les filtres gammatones. On démontre aussi, que par un placement approprié des zéros, les propriétés physiologiques des réponses impulsionnelles du nerf auditif peuvent être reproduites fidèlement.

2.1.2 Dérivation des filtres binomiaux

La loi binomiale de paramètres m et p est la loi de probabilité d'une variable aléatoire X égale au nombre de succès rencontrés au cours d'une répétition de m épreuves de Bernoulli, p étant la probabilité de succès d'une épreuve de Bernoulli. Avoir exactement k $(0 \leq k \leq m)$ succès durant cette experience a une probabilité donnée par P_m^k :

$$P_m^k = \binom{m}{k} p^k (1-p)^{m-k} \tag{2.9}$$

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} \tag{2.10}$$

Si t représente la variable temps $(t \geq 0)$, il existe un λ positif tel que $p = \exp(-\lambda t)$. Dans ce cas, p représente la probabilité instantanée d'un succès d'une épreuve de Bernoulli. Si m réalisations indépendantes du même processus se produisent simultanément, la probabilité d'avoir exactement k sucées à n'importe quel instant $t \geq 0$ est donnée par $P_m^k(t)$:

$$P_m^k(t) = \binom{m}{k} \exp(-k\lambda t)[1 - \exp(-\lambda t)]^{m-k} \tag{2.11}$$

Dans figure 2.1, $P_m^k(t)$ est tracée pour différentes valeurs de k et λ. Ces valeurs définissent la forme de cette distribution. Par exemple, la valeur maximale est atteinte pour $t_{max} = log(m/k)/\lambda$ alors que k contrôle la valeur de la pente au voisinage de 0.

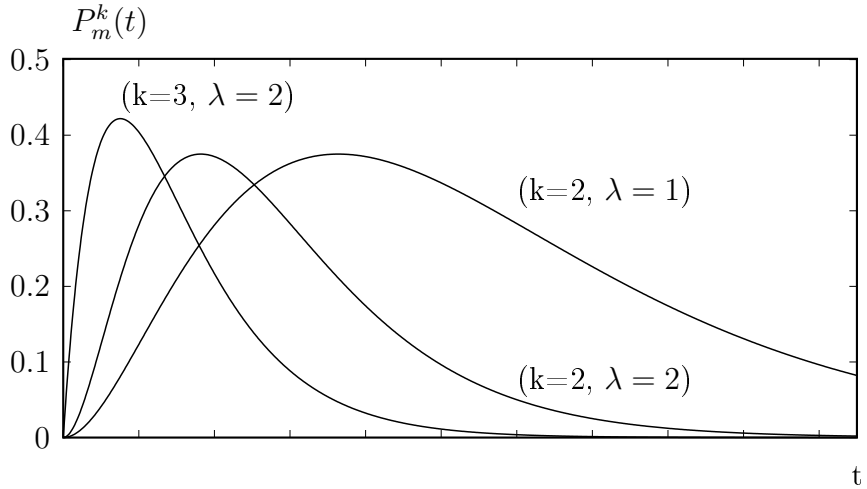


Figure 2.1 Distribution binomiale pour différentes valeurs de ses paramètres ($m = 4$, λ est un réel positif, t est donné en unité arbitraire).

On définit le filtre binomial, *Binomial-tone filter* (BIT) par sa réponse impulsionnelle donnée par :

$$\mathrm{BITF}_m^k(t) = A\exp(-\lambda kt)[1-\exp(-\lambda t)]^{m-k}\cos(\omega_0 t+\phi) \tag{2.12}$$

Où ω_0 est la fréquence de résonance, ϕ est la phase initiale et A est une constante.

Discussions : Les filtres BITs (*Binomial-tone filter*) présentés dans cette thèse ne sont pas à confondre avec les filtres gaussiens. Ces derniers sont des filtres passe-bas utilisés pour le filtrage des images par exemple (*gaussian blur filter*) [Aubury et Luk, 1996; Haddad et Akansu, 1991]. Une approche efficace pour implémenter un filtre gaussien sous forme d'un filtre à réponse impulsionnelle finie (RIF) consiste à approximer la distribution normale en utilisant les coefficients binomaux (voir l'équation (2.10)). En effet, on montre que la convolution du filtre $[1,\ z^{-1}]$ N fois :

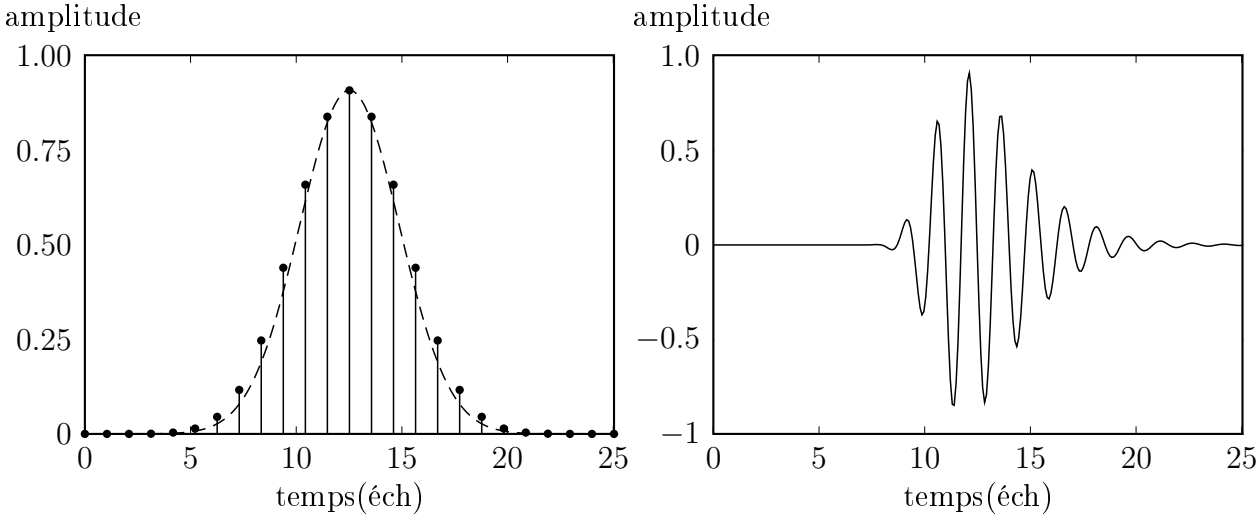
$$B_N = \underbrace{[1,\ z^{-1}] * [1,\ z^{-1}] * \cdots * [1,\ z^{-1}]}_{\text{N fois}} \tag{2.13}$$

forme une bonne approximation de la RI du filtre gaussien donnée par l'équation (2.14) [Crowley *et coll.*, 2002].

$$g(m) = \frac{1}{\sigma\sqrt{2\pi}} e^{m^2/2\sigma^2} \tag{2.14}$$

Où $m = n - N/2$ et $\sigma = \sqrt{N}/2$. La figure 2.2 illustre les différences entre la RI du filtre gaussien et celles des filtres binomiaux présentés dans cette thèse. Même si ces deux filtres font référence au même terme (à savoir *binomial*) dans leur nom respectif, leurs expressions et leurs applications sont fondamentalement différentes. On résume dans ce qui suit les différences majeures entre les filtres gaussiens et les filtres BITs présentés dans cette thèse :

- Les filtres gaussiens sont des filtres à réponse impulsionnelle finie alors que les filtres BITs sont des filtres à réponse impulsionnelle infinie.
- Les filtres gaussiens sont des filtres passe-bas alors que les filtres BITs sont des filtres passe-bande.
- Les coefficients des filtres gaussiens sont tous positifs alors que ceux des filtres binomiaux présentés dans cette thèse ont des coefficients positifs et négatifs (voir équation (2.16)).
- L'enveloppe des RIs des filtres gaussiens est symétrique alors que celle des BITs ne l'est pas (voir la figure 2.1 ou la figure 2.2(b)).



(a) Réponse impulsionnelle du filtre gaussien (ligne discontinue) approximée par les coefficients binomiaux (symbole •).

(b) Réponse impulsionnelle du filtre binomial proposé.

Figure 2.2 Comparaison entre les filtres gaussiens et les filtres proposés.

2.1.3 Implementation digitale des filtres binomiaux

La transformée en z ($\mathcal{Z}$) ainsi que celle de Laplace ($\mathcal{L}$) du filtre BIT peuvent être calculées facilement. Il suffit de remarquer que le terme $[1-\exp(-\lambda t)]^{m-k}$ se développe en une somme pondérée en utilisant la formule du binôme de Newton. On peut alors écrire :

$$\begin{aligned} H_{\mathrm{BITF}_n^k}(z) &= A \times \mathcal{Z}\Big[[1-\exp(-\lambda t)]^{m-k}\left[\exp(-\lambda k t)\cos(\omega_0 t+\phi)\right]\Big] \\ &= A \times \sum_{l=0}^{n}\binom{n}{l} \times \mathcal{Z}\left[(-\beta^{lt}) \times \mathcal{Z}^{-1}\left(\frac{a(z)}{b(z)}\right)\right] \end{aligned}$$

La transformée en $\mathcal{Z}$ d'une sinusoïde amortie est donnée par [Healey, 1967] :

$$\mathcal{Z}\left(\exp(-\lambda k t)\cos(\omega_0 t+\phi)\right) = \frac{\cos(\phi)z^{-1} - e^{-k\lambda}\cos(w_0-\phi)z^{-2}}{1-2e^{-k\lambda}\cos(w_0)+e^{-2k\lambda}z^{-2}} = \frac{a(z)}{b(z)};$$

$$\beta = \ln(\lambda); n = m-k$$

En utilisant la propriété du *scaling*[2] de la transformée en z [Tohyama et Koike, 1998], $H_{\mathrm{BITF}_n^k}(z)$ est donnée par :

$$H_{\mathrm{BITF}_n^k}(z) = \frac{AZ(z)}{BZ(z)} = \frac{\sum_{l=0}^{n} C_n^l a_l(z) \prod_{\substack{j=0 \\ j\neq l}}^{n} b_j(z)}{\prod_{l=0}^{n} b_l(z)} \tag{2.15}$$

Où :

$$C_n^l = \binom{n}{l}(-1)^l \tag{2.16}$$
$$a_l(z) = a(\beta^{-l}z);\ b_l(z) = b(\beta^{-l}z)$$

Puisque $\sum C_n^l = 0$ alors AZ(1)=0 et le nombre total des coefficients de $H_{\mathrm{BITF}_n^k}(z)$ est donné par $4n+5$ pour un ordre $n \geq 1$ donné.

La même procédure peut être utilisée pour trouver la transformée de Laplace du filtre BIT. Cette dernière est donnée par :

$$H_{\mathrm{BITF}_n^k}(s) = \sum_{l=0}^{n} \binom{n}{l} \frac{\cos(\phi)(s + s_0 + l\beta)}{(s + p_0 + l\beta)^2 + \omega_0^2} \tag{2.17}$$

Avec $p_0 = \lambda k$ et $s_0 = -\tan(\phi)\omega_0 + \lambda k$.

Sur la figure 2.3, le spectre du filtre BIT est donné pour les mêmes valeurs que celles utilisées pour la figure 2.1. Les deux paramètres λ et n changent la valeur du facteur de qualité du filtre BIT : pour des valeurs élevées de n ou λ le facteur d'amortissement est plus élevé d'où un spectre plus plat.

La phase initiale ϕ introduit une certaine asymétrie sur le spectre mais ce dernier reste relativement symétrique pour être un modèle approprié des RIs du NA. Le spectre du filtre GT de quatrième ordre est donné aussi sur la même figure. Au voisinage de la fréquence de résonance, le spectre du filtre BIT ressemble à celui du GT mais la différence est plus prononcée au niveau des basses et hautes fréquences. Dans la section 2.4, on donne plus de détails concernant la relation entre le filtre BIT et le filtre GT.

2. Si $\mathcal{Z}(f(t)) = F(z)$, alors $\mathcal{Z}(\lambda^t f(t)) = F(\lambda^{-1}z)$

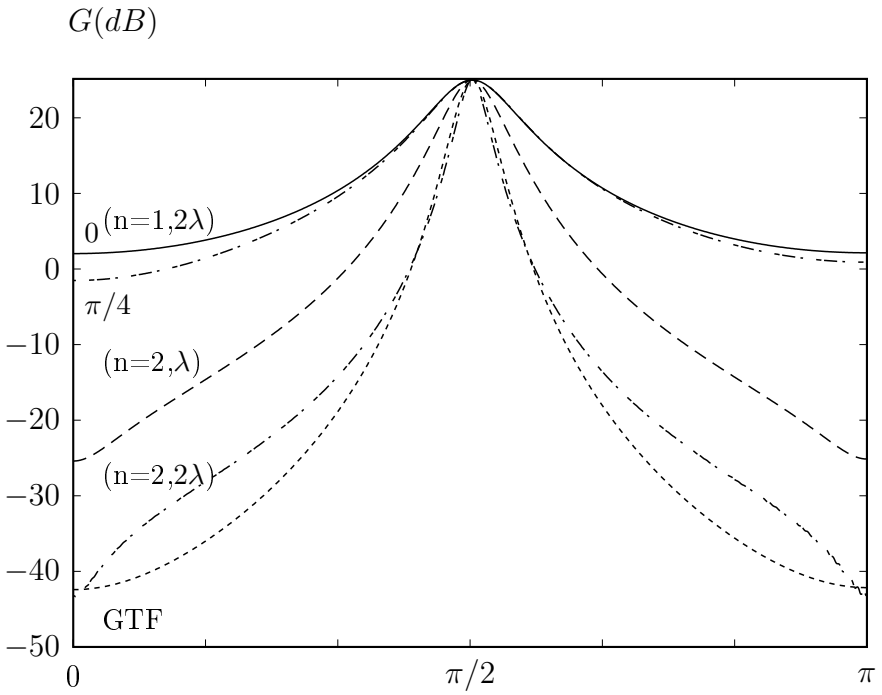


Figure 2.3 Spectre du filtre binomial pour différentes valeurs des paramètres ($n = 2$, λ est une valeur positive et $\phi = 0, \pi/4$). Le spectre du GTF de 4ième ordre est tracé en pointillé.

Le filtre BIT partage avec le filtre GT les mêmes limitations : spectre symétrique, trajectoire des FIs constante et absence de compression au voisinage de la MF. Dans la section suivante, on propose des modifications dont le but est d'adapter le filtre BIT aux observations physiologiques. Les modifications proposées n'introduisent pas de complexité additionnelle en terme d'implémentation.

2.2 Adaptation des filtres binomiaux aux observations physiologiques

2.2.1 Réponses impulsionnelles du nerf auditif

Dans [Carney *et coll.*, 1999], la réponse impulsionnelle du nerf auditif d'une population de chats a été mesurée en utilisant la technique de la corrélation-inversée et ce pour différentes fréquences et différents niveaux d'excitation. On suit la même procédure que celle utilisée par [Irino et Patterson, 2001] et ce dans le but de fournir des résultats comparables à

ceux fournis pour les filtres GT et les filtres GC. On utilise aussi les mêmes réponses impulsionnelles qui sont disponibles sur le site web du earLab [3].

La méthode utilisée pour collecter ces réponses est détaillée dans la section I de [Carney *et coll.*, 1999]. Les réponses impulsionnelles de la cellule 25 de l'Unité 86100 de MF de 2 kHz seront utilisées pour le reste de ce chapitre puisque c'est l'unité avec le plus grand nombre de RIs disponibles.

Enveloppe temporelle

Sur la figure 2.4, la RI de l'Unité 86100u25 est donnée pour le niveau d'excitation de 80 dB SPL. Les enveloppes des mêmes RIs sont données sur la figure 2.5. L'enveloppe

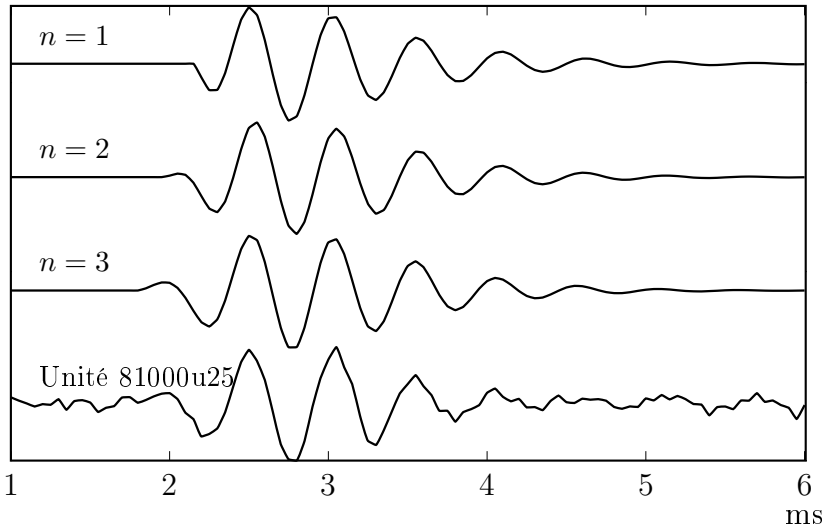


Figure 2.4 Réponse impulsionnelle du filtre binomial pour différents ordres du modèle. La réponse impulsionnelle de l'Unité 86100u25 de [Carney *et coll.*, 1999] est aussi donnée (80 dB SPL).

est déterminée comme étant la valeur absolue de la transformée de Hilbert ($\mathcal{H}$) de ces réponses impulsionnelles. La différence entre la RI du modèle et celle du NA est plus prononcée au début de cette dernière. Un modèle dont l'ordre est plus élevé correspond à une erreur plus petite. Dans la section 2.2.2 on donnera plus de détails sur le lien entre l'ordre du modèle et l'erreur de modélisation. Il parait à premiere vue que l'enveloppe du filtre BIT fournit une bonne approximation de l'enveloppe du NA. La propriété du NA la plus difficile à modéliser est le taux de glissement des FIs qui se trouve être indépendant du niveau d'excitation. On définit le taux de glissement des RIs comme étant la pente de la variation de la fréquence instantanée de ses dernières en fonction du temps.

3. `http://earlab.bu.edu/databases/collections/Default.aspx`

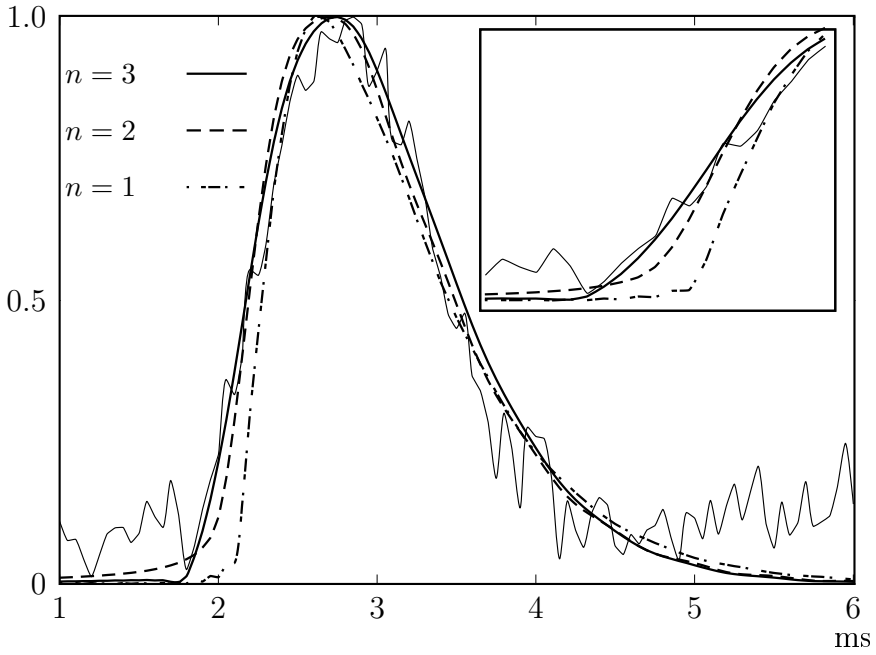


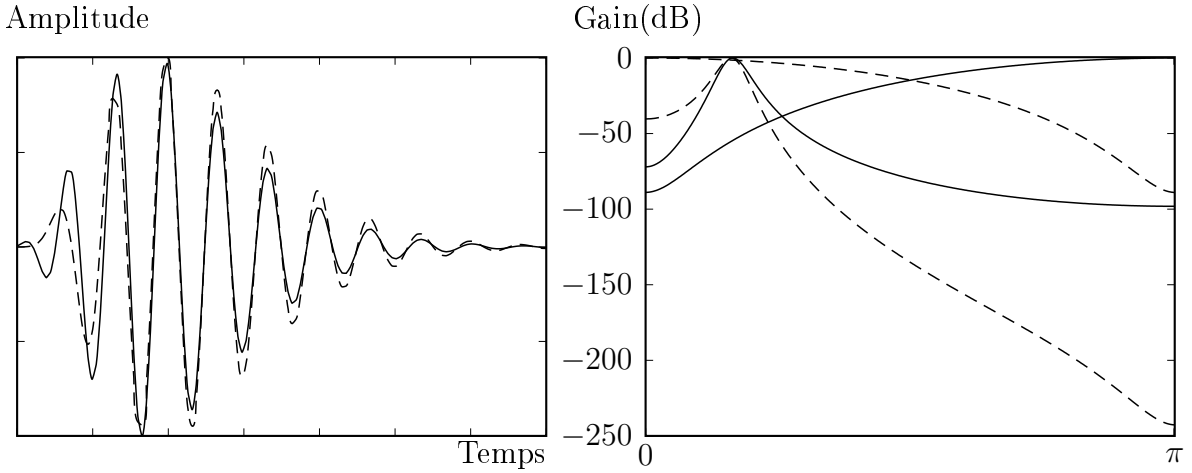
Figure 2.5 Enveloppe estimée à partir de la réponse impulsionnelle de l'Unité 81000u25 [Carney *et coll.*, 1999] (80 dB SPL). L'enveloppe des réponses impulsionnelles du BITF est donnée pour différents ordres.

Fréquences instantanées et asymétrie spectrale

Les FIs du NA présentent des taux de glissement dont la direction dépend seulement de la MF de la fibre auditive. Dans le cas du filtre BIT, le changement de la phase initiale permet de changer le taux de glissement de la fréquence instantanée (figure 2.3). Mais généralement, le spectre du filtre BIT tel que décrit dans l'équation (2.15) est symétrique et la modification de la phase initiale ϕ ne permet pas d'introduire une fréquence instantanée dont la trajectoire est controllable.

On propose dans ce qui suit des modifications de l'équation (2.15) dans le but d'introduire un taux de glissement controllable dans les fréquences instantanées du filtre BIT. Ceci peut être effectué facilement en ajoutant un nouveau paramètre γ qui contrôle la position des zéros dans l'équation (2.15). En effet, il suffit de remplacer $a_k(z)$ dans (2.15) avec $a_k(z) = a(\gamma^{-k}z)$. En décalant $AZ(z)$ par une valeur de π, ce dernier peut être un filtre passe-haut ou un passe-bas ce qui permet d'introduire un glissement des FIs dont la trajectoire est controllable.

Sur la figure 2.6 deux RIs du filtre BIT sont présentées. Ces deux réponses partagent la même enveloppe et la même FI asymptotique. Cependant, au début de la RI les trajectoires



(a) Réponses impulsionnelles du filtre binomial pour $n = 2$

(b) Spectres du filtre binomial pour $n = 2$

Figure 2.6 Spectres et réponses impulsionnelles du filtre binomial dont les RIs ont des fréquences instantanées dont le taux de glissement est décroissant (ligne en pointillé) et croissant (ligne continue). Sur la figure 2.6(b) les réponses du filtre $AZ(z)$ ainsi que celles du filtre composite $AZ(z)/BZ(z)$ (équation (2.15)) sont données pour chaque cas.

sont différentes. La RI représentée en pointillé sur la figure 2.6(a) possède un taux de glissement croissant alors que celle en ligne continue un taux de glissement décroissant. Cette différence est aussi visible sur leurs spectres. Les deux spectres donnés sur la figure 2.6(b), ont les mêmes pôles ($BZ(z)$ dans l'équation (2.15)) mais des zéros différents.

Compression et déviation fréquentielle

Parmi les manifestations du comportement non-linéaire de la cochlée, on trouve la compression des niveaux sonores [Allen, 2001], la suppression due à la présence d'une seconde tonalité [Ruggero *et coll.*, 1997] et la distorsion des produits oto-émissions acoustiques [Cooper et Rhode, 1997]. L'un des phénomènes non linéaires des plus importants est la compression des niveaux sonores élevés. Les signaux de faible intensité sont amplifiés avec des gains élevés, alors que ceux de hauts niveaux ne sont pratiquement pas amplifiés. Ainsi la cochlée présente une compression croissante en fonction de l'intensité du signal d'entrée : la cochlée réalise un contrôle de gain automatique de telle sorte que son gain devient atténué pour des signaux dont l'intensité croît. La figure 2.7 présente le spectre des RIs de la cellule 25 de l'Unité 86100 dont la MF est de 2 kHz. Le gain de ses RIs au voisinage de la fréquence centrale est inversement proportionnel au niveau d'excitation du signal d'entrée. Ces spectres présentent aussi une asymétrie (indépendante du niveau d'excitation) qui suppose un taux de glissement constant des FIs. La direction pointée par

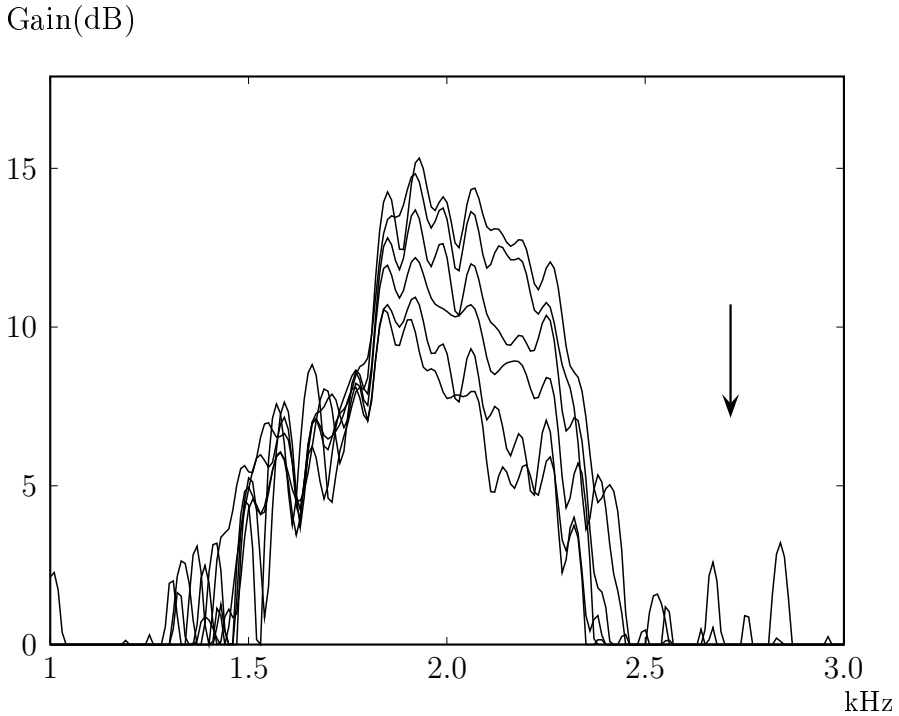


Figure 2.7 Spectres des réponses impulsionnelles des cellules de l'Unité 86100u25. La direction pointée par la flèche indique des niveaux d'excitation croissants. Quand le niveaux d'excitation augmente, la valeur maximale des spectres des RIs diminuent d'où la compression au voisinage de la fréquence centrale.

la flèche indique des niveaux d'excitation croissants. Quand le niveaux d'excitation augmente, la valeur maximale des spectres des RIs diminue d'où la compression au voisinage de la fréquence centrale.

On propose dans ce qui suit de valider le filtre BIT par rapport aux RIs du nerf auditif de chat et ce en ce qui concerne :

- L'allure temporelle de la réponse impulsionnelle.
- Le taux de glissement de la FI qui est indépendant du niveau d'excitation.
- La déviation de la MF en fonction du niveau d'excitation.
- La compression au voisinage de la fréquence centrale.

2.2.2 Modélisation des réponses impulsionnelles du nerf auditif

On étudie dans cette section la validité du filtre BIT comme un modèle de la réponse impulsionnelle du nerf auditif. La variation des paramètres du modèle est exprimée en fonction du niveau d'excitation utilisée pour recueillir les réponses impulsionnelles publiée par [Carney *et coll.*, 1999]. La réponse du modèle est ensuite comparée à celle recueillie au niveau du nerf auditif. On propose donc dans cette section de trouver les valeurs des paramètres du BIT dont les RIs ressemblent le mieux à celles du NA.

Modèle et erreur de modélisation

Les paramètres du filtre BIT sont ajustés directement dans le domaine de la transformée en z et sont implémentés comme décrit par l'equation (2.15). Le filtre biquadratique $a(z)/b(z)$ est paramétré dans le domaine fréquentiel où G, w_c, w_l et w_r représentent le gain maximal, la fréquence centrale, les basses fréquences et hautes fréquences à -3dB respectivement. La fréquence d'échantillonnage est de 20kHz.

Pour un ordre du modèle n donné, les paramètres qui sont impliqués dans l'ajustement du filtre BIT sont le gain maximal G, β, γ et les paramètres fréquentiels $\delta_c = w_c/MF$, $\delta_r = w_r/MF$ and $\delta_l = w_l/MF$. La variation de ces paramètres en fonction du niveau d'excitation P_N est exprimée comme étant un système linéaire donné par :

$$\begin{pmatrix} G \\ \beta \\ \gamma \\ \delta_c \\ \delta_l \\ \delta_r \end{pmatrix} = \mathbf{f}(P_x) = \mathbf{M} \times P_x + b_0 \tag{2.18}$$

Où $M \in \mathbb{R}_{6,1}$ et P_x est le niveau d'excitation normalisé (exprimé en dB SPL) donné par $(P_N - 80)$ [4].

Algorithme d'apprentissage

La RI du modèle ainsi que celle du NA de la population de chats sont normalisées puisque les gains des RIs de chats vers les basses fréquences n'ont pratiquement pas changés quand le niveau d'excitation a changé [Carney *et coll.*, 1999] (voir figure 2.7). Le filtre BIT est utilisé avec 6 paramètres dont la variation est modélisée par un système linéaire décrit par l'équation (2.18). Le délai des RIs n'est pas investigué dans cette étude et est déterminé comme étant le délai pour lequel la corrélation entre la RI du filtre BIT et celle des RIs des cellules auditives de chats atteint son maximum. L'erreur quadratique moyenne (EQM) est utilisée comme critère de minimisation pour la procédure d'apprentissage du modèle. L'erreur d'apprentissage est définie comme étant l'EQM de la différence entre la réponse du modèle et celle de la RI du NA. Cette différence est calculée sur la durée pour laquelle l'enveloppe des RIs est supérieure à une fois et demie (1.5) le bruit de mesure estimé à partir du premier et des deux dernières ms de la RI du NA.

Si un modèle $\mathbf{f}$ (conformément à l'équation (2.18)) est utilisé pour générer des RIs notées $\hat{RI}_{\mathbf{f}}(i)$, l'erreur $e(\mathbf{f})$ est définie par :

$$e(\mathbf{f}) = \frac{\sum_{i=1}^{N} \left(RI(i) - \hat{RI}_f(i) \right)^2}{\sum_{i=1}^{N} RI(i)^2} \tag{2.19}$$

Où N représente le nombre des RIs par fibre auditive.

La valeur finale de l'EQM est ensuite normalisée par la puissance moyenne de la RI du NA et exprimée en dB.

L'algorithme de minimisation est un algorithme de recherche linéaire itératif basé sur l'algorithme de la plus forte pente où l'estimation du Jacobien (∇) est effectuée numériquement. À chaque itération, la valeur de $\mathbf{f}$ est mise à jour en utilisant l'algorithme 2.1.

4. Ce choix est motivé par le fait que le niveau d'excitation commun aux réponses impulsionnelles est de 80 dB SPL.

Algorithme 2.1 : Algorithme d'ajustement du BIT aux RIs du NA

Entrées : P_N : Niveaux d'excitation sonore.
Sorties : $\mathbf{f}$ définie dans (2.18).
Données : MF : fréquence centrale des RIs
tant que $|\nabla e(\mathbf{f})| \geq \epsilon$ **faire**
  Générer les RIs du BIT en utilisant $\mathbf{f}$.
  Estimer $\Delta e(\mathbf{f})$, le gradient de $e(\mathbf{f})$.
  Trouver α pour minimiser $e(\mathbf{f} + \alpha \Delta e(\mathbf{f}))$.
  Mettre à jour $\mathbf{f} = \mathbf{f} - \alpha \Delta e(\mathbf{f})$.
fin

On présente dans la section suivante les résultats obtenus suite à l'ajustement du modèle aux RIs du NA et on démontre que le filtre BIT est un modèle à complexité réduite capable de modéliser fidèlement les RIs du NA.

2.3 Résultats expérimentaux

2.3.1 L'ordre du modèle

Dans cette section, on examine l'impact de l'ordre du modèle n sur l'erreur de modélisation. Pour ce faire, les RIs générées par le filtre BIT sont ajustées aux RIs du NA et ce que pour des niveaux d'excitation sonores de 80 dB SPL. L'ajustement du modèle est limité dans ce cas aux niveaux d'excitation P_N élevés et ce pour deux raisons :

1. Pour des niveaux d'excitation élevés, le bruit de mesure est le plus faible.
2. Ajuster le modèle à une RI par fibre auditive, évite l'impact du nombre de coefficients du modèle sur l'erreur de modélisation (section 2.3.2).

La figure 2.8 représente l'erreur de modélisation calculée selon l'équation (2.19) exprimée en dB.

Quand le modèle est un simple filtre biquadratique ($n = 0$), l'erreur de modélisation est la plus élevée et vaut -6.3 dB. Quand l'ordre du modèle augmente, l'erreur de modélisation diminue ensuite augmente légèrement à partir de $n \geq 5$. Ceci est dû essentiellement au fait que pour $n \geq 5$, le terme $(1 - \beta^t)^n$ ne peut croître aussi rapidement que l'accroissement de l'enveloppe les RIs du NA.

Étant donné qu'un ordre de modélisation plus élevé implique une implémentation digitale avec un nombre de coefficients plus élevé, un compromis raisonnable semble être possible avec un ordre de modélisation valant 2 ou 3.

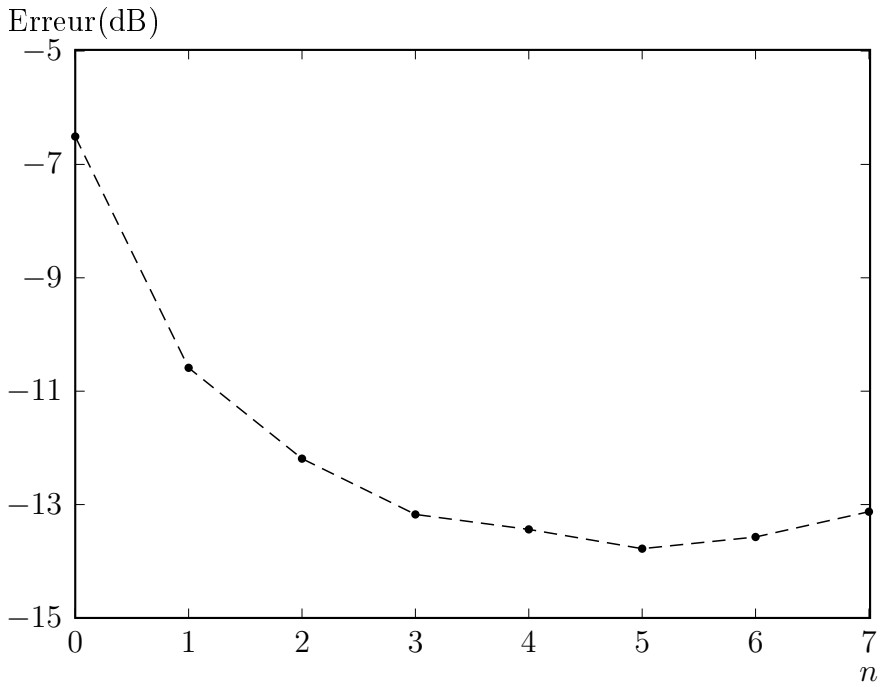


Figure 2.8 Compromis entre ordre du filtre binomial (n) et erreur de modélisation des réponses impulsionnelles du nerf auditif.

2.3.2 Résultats de modélisation de la fibre 25 de l'Unité 86100

Le nombre des coefficients du modèle linéaire f

On s'intéresse dans cette section à l'impact du nombre des coefficients du modèle $\mathbf{f}$ sur l'erreur de modélisation : le nombre maximal par paramètre est 2 (ordonnée à l'origine $\mathbf{b_0}$ et coefficient directeur $\mathbf{M}$) et le nombre total des coefficients de $\mathbf{f}$ est de 12 (6 paramètres équation (2.18)). Puisque le taux de glissement des FIs est indépendant du niveau d'excitation P_N, le paramètre γ contrôlant le taux de glissement des RIs du modèle $\mathbf{f}$ est modélisé par une constante.

Dans le tableau 2.1, le nombre total des coefficients est donné dans la première colonne. La distribution de ces coefficients parmi les différents paramètres est aussi donnée. Par exemple, sur la deuxième ligne, 10 coefficients sont utilisés : tous les paramètres du modèle dépendent du niveau d'excitation sauf les deux paramètres γ et β. L'impact de l'ordre de modélisation n sur l'erreur de modélisation e est aussi investigué.

Le tableau 2.1 montre que le filtre BIT peut modéliser fidèlement les RIs des fibres auditives étant donné que le rapport signal sur bruit (RSB) de ces dernières est d'environ 20 dB. Pour le même nombre de coefficients un modèle dont l'ordre est plus élevé résulte en une erreur de modélisation plus petite (une réduction de 0.3 dB) grace à une meilleure modélisation

Tableau 2.1 Compromis entre l'ordre du modèle n, le nombre de coefficients du modèle et l'erreur de modélisation des réponses impulsionnelles de la fibre 25 de l'Unité 86100 pour 9 niveaux d'excitation.

						n=2		n=3	
Nombre de coeff.	G	β	w_c	δ_l	δ_r	e	cmp	e	cmp
11	2	2	2	2	2	-13.8	0.76	-13.4	0.70
10*	2	1	2	2	2	-13.2	0.81	-13.3	0.56
9	2	1	2	1	2	-12.9	0.73	-13.1	1.00
8	2	2	1	1	1	-11.7	0.81	-11.8	1.00
8	2	1	2	1	1	-12.2	0.82	-12.7	0.99
8	2	1	1	1	2	-12.9	0.78	-12.3	0.38
7	2	1	1	1	1	-11.5	0.81	-11.7	0.87
6	1	1	1	1	1	-10.9	1.00	-11.1	1.00

du début des RIs. Quand l'ordre du modèle est fixé, cette même erreur peut être réalisée avec un nombre de coefficients plus petit : Il suffit d'affecter deux coefficients pour δ_r et w_c. En effet, le paramètre δ_r contrôle la pente descendante de l'enveloppe de la RI du filtre BIT ce qui est consistent avec le fait que la durée des RIs décroît quand le niveau d'excitation augmente. Le paramètre w_c doit aussi dépendre du niveau d'excitation P_N puisque le pic du spectre des RIs du NA se décale vers les basses fréquences quand le niveau d'excitation augmente (Figure 2.7). Quand les paramètres du modèle sont indépendants du niveau d'excitation l'erreur de modélisation est la plus grande et vaut -10.9 dB.

Clairement, les paramètres du filtre BIT doivent être dépendants du niveau d'excitation. Selon le tableau 2.1, les paramètres qui influencent le plus l'erreur de modélisation sont δ_r (lignes 6 et 7 du tableau 2.1) et w_c. Quand ces deux paramètres sont dépendants du niveaux d'excitation, l'erreur de modélisation est inférieure à -12.9 dB. La dépendance du paramètre δ_l améliore légèrement l'erreur de modélisation avec environ 0.3 dB (lignes 2 et 3 du tableau 2.1) alors que celle du paramètre β est seulement importante quand le paramètre δ_r est lui aussi dépendant du niveau d'excitation (lignes 1 et 2 du tableau 2.1). Même si le paramètre β influence peu l'erreur de modélisation, il influence grandement le comportement compressif du filtre BIT. En effet, ce paramètre contrôle l'énergie de la RI du filtre BIT et modéliser ce paramètre comme étant dépendant du niveau d'excitation mène à un modèle surdéterminé (10 paramètres pour 9 RIs du NA).

Compression au voisinage de la fréquence centrale

Le taux de compression τ_{cmp} est défini comme étant la pente qui décrit la variation moyenne du gain maximal du spectre de la RI par rapport à la variation du niveau d'excitation P_N.

$$\tau_{\text{cmp}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta\text{MAX(TFD[RI(i)])}}{\Delta P_N} \tag{2.20}$$

Où TFD représente la transformée de Fourier discrète (TFD) et N représente le nombre des RIs du NA.

Pour la fibre 25 de l'Unité 86100 [Carney *et coll.*, 1999], il semble que le modèle se comporte relativement pareillement pour $n = 2$ et $n = 3$, sauf quand il s'agit de compression. En effet, pour $n = 3$ et quand tous les paramètres du modèle dépendent du niveau d'excitation, le taux de compression moyen τ_{cmp}=0.56 dB/dB ce qui est consistent avec les valeurs de compression de la MB des cochons d'inde [Cooper et Yates, 1994](pour des fréquence d'excitation de 2kHz). Cette valeur est aussi consistante avec celle dérivée des experiences de masquage fréquentiel [Hicks et Bacon, 1999].

Les paramètres qui contrôlent la compression au niveau de la crête du filtre BIT sont β et δ_r. En effet, β est responsable de la pente d'accroissement de l'enveloppe du filtre BIT alors que δ_r contrôle le décroissement de la même enveloppe. Permettre à ces deux paramètres d'être dépendants du niveau d'excitation, garanti que l'enveloppe du filtre BIT coincide parfaitement avec celle des RIs du NA. Quand la durée des RIs du NA augmente (niveau d'excitation décroît), le spectre devient plus plat et donc le gain maximal du filtre BIT décroît d'où la compression.

L'implémentation digitale du filtre BIT requière $4n + 5$ coefficients, donc pour $n = 3$ le nombre de coefficients est 17 et 13 pour le cas de $n = 2$. Une autre façon d'introduire la compression consiste à ajouter un pôle additionnel au filtre BIT d'ordre 2. Ceci consiste à modifier l'équation (2.15) comme suit :

$$H_{\text{CBIT}_n^k}(z) = z^{-m}\frac{\displaystyle\sum_{k=0}^{n} C_n^k a_k(z)\prod_{\substack{i=0\\ i\neq k}}^{n} b_i(z)}{\displaystyle\prod_{k=0}^{n+1} b_k(z)} \tag{2.21}$$

L'équation (2.21) définit la fonction de transfert du filtre binomial compressif, *Compressive Binonial-tone filter* (cBIT) dont l'implémentation digitale nécessite seulement 15 coefficients pour $n = 2$.

Le modèle implémenté tel que décrit par l'équation (2.21) est ajusté aux RIs de la fibre 25 du NA de l'Unité 86100 et ce pour 9 niveaux d'excitation en utilisant la relation (2.18).

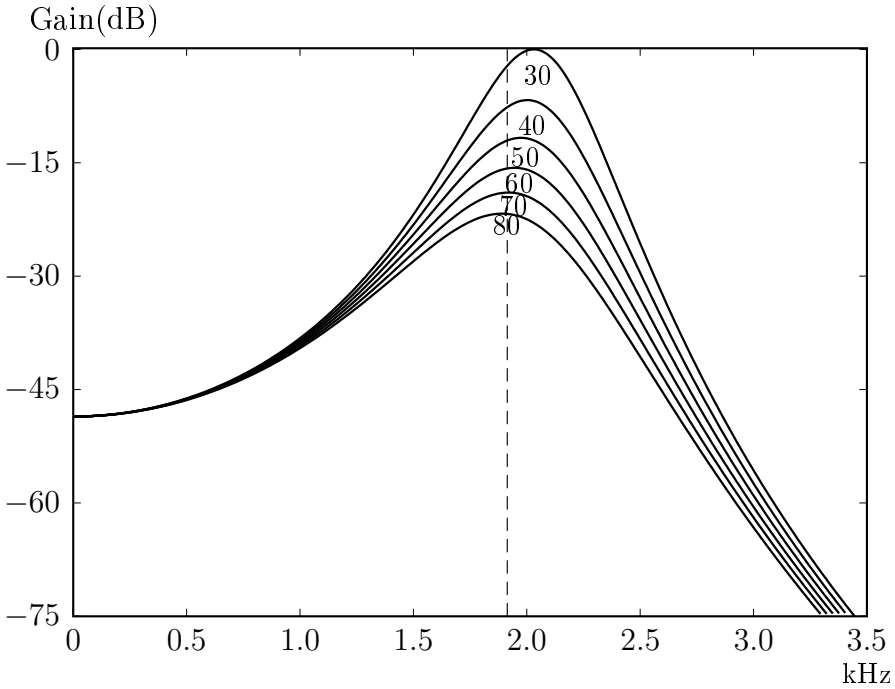
Résultats de modélisation

Les résultats de modélisation de l'Unité 86100u25 sont présentés sur la figure 2.9. La figure 2.9(a) présente le spectre du filtre cBIT dont les RIs sont présentées sur la figure 2.10(a). La fréquence de résonance w_c du filtre cBIT diminue de façon monotone avec le niveau d'excitation et est à peu près égale à la MF (1.9kHz représenté en pointillé sur la figure 2.9(a)) pour un niveau d'excitation de 80 dB SPL. Le gain du filtre cBIT varie aussi avec le niveau d'excitation. Sur la figure 2.9(a), le gain a été normalisé à 0 dB pour un niveau de 30 dB SPL. Le gain diminue avec le niveau d'excitation et est d'environ -21 dB pour un niveau d'excitation de 80 dB SPL. Et donc, la fonction d'entrée-sortie du filtre est compressive et le taux de croissance moyen est d'environ 0.57 dB/dB. Ces filtres ont également un spectre stable pour les basses fréquences ce qui est compatible avec le fait que les trajectoires des FIs sont indépendantes du niveau d'excitation. Cela peut aussi se voir sur les réponses impulsionnelles (figure 2.10) qui partagent les mêmes temps de passage par zéro que les RIs du NA.

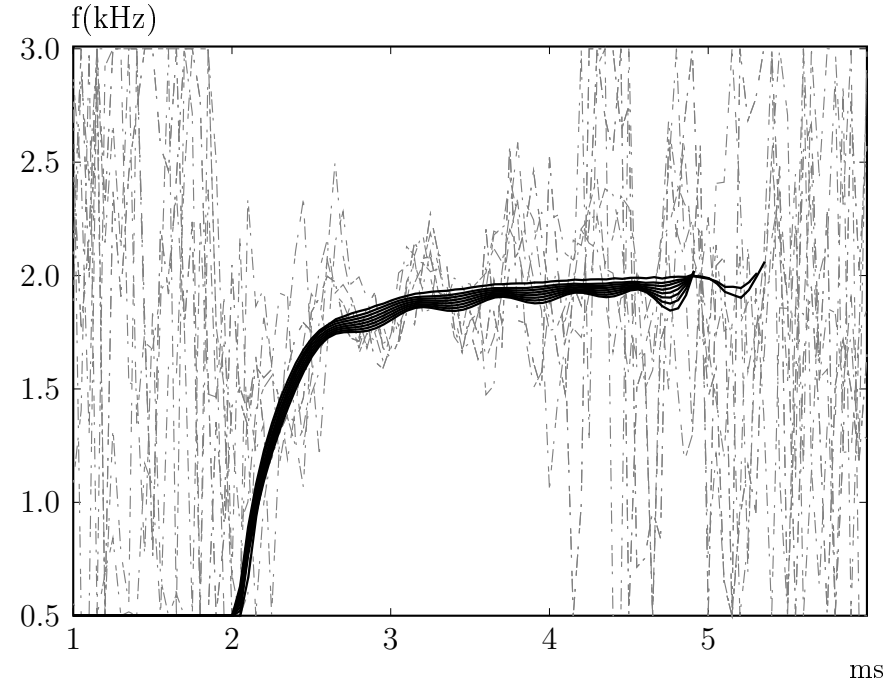
L'évolution temporelle de l'erreur entre les RIs du NA et les réponses impulsionnelles du filtre cBIT est donnée sur la figure 2.10(b). Les disparités entre ces dernières sont plus prononcées aux parties finales des RIs à 30 et 40 dB : les RIs du NA et celle du filtre cBIT sont presque en coïncidence lorsque l'amplitude relative de la forme d'onde est suffisamment grande. Ceci est probablement dû au fait que les paramètres du modèle ont été modifiés de façon linéaire par rapport au niveau d'excitation : un ordre de modélisation plus élevé (polynôme du second ordre) aurait donné de meilleurs résultats.

Indépendamment de la dernière partie des RIs du NA, le début de ces dernières est bien modélisé et les temps de passage par zéro sont conservés. Ceci est également illustré sur la figure 2.9(b) : Les FIs du filtre cBIT montent doucement en suivant la moyenne des FIs des RIs du NA sur la durée durant laquelle le rapport signal sur bruit est élevé.

La figure 2.11 montre le déplacement des pôles et zéros du filtre cBIT quand le niveau d'excitation varie. Le filtre cBIT d'ordre 2 possède 4 pôles, leurs conjugués et 2 zéros purement réels de multiplicité 2 chacun. Lorsque le niveau d'excitation augmente, les 4 pôles ainsi que leurs conjugués migrent vers le centre du cercle unité. Cela a deux

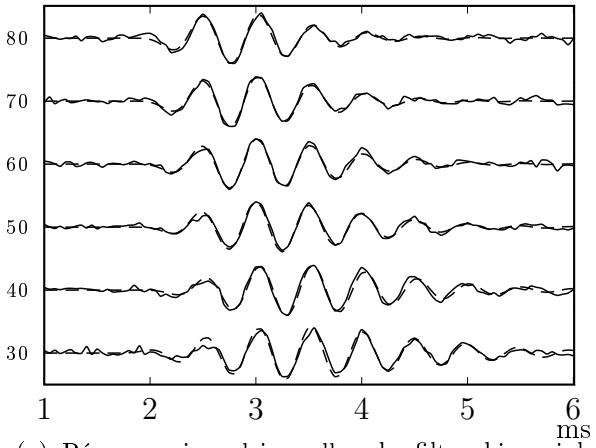


(a) Spectres du filtre binomial compressif pour 6 niveaux d'excitation. Le niveau d'excitation est donné sur la même figure.

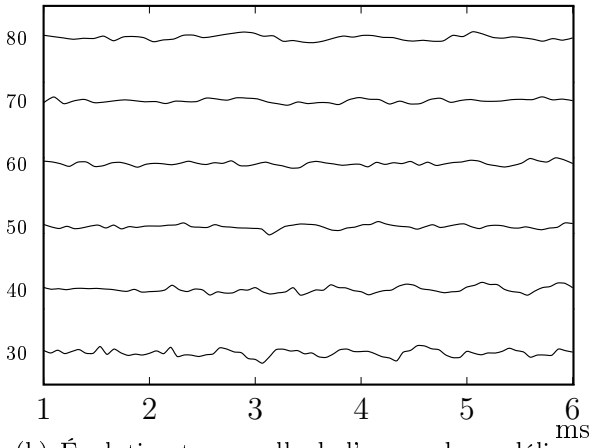


(b) Trajectoires des fréquences instantanées du filtre binomial compressif (ligne solide) et celles de la fibre auditive 25 (pointillé).

Figure 2.9 Spectres et trajectoires des fréquences instantanées du filtre binomial compressif pour l'Unité 86100u25.

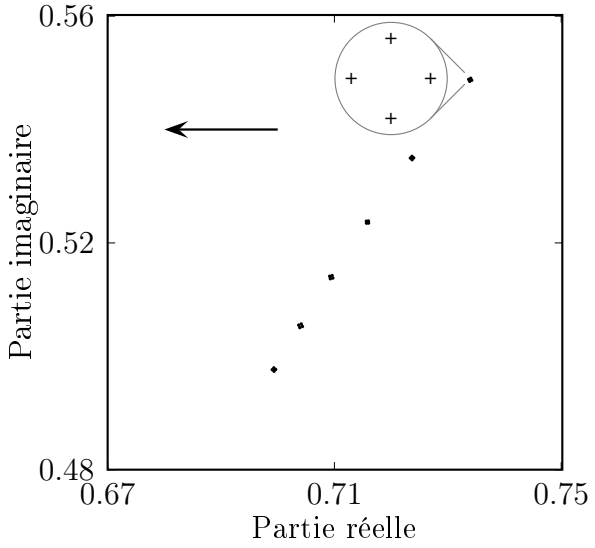


(a) Réponses impulsionnelles du filtre binomial compressif (ligne pointillée) et celles du nerf auditif. Le niveau d'excitation en dB est donné sur l'axe des ordonnées.

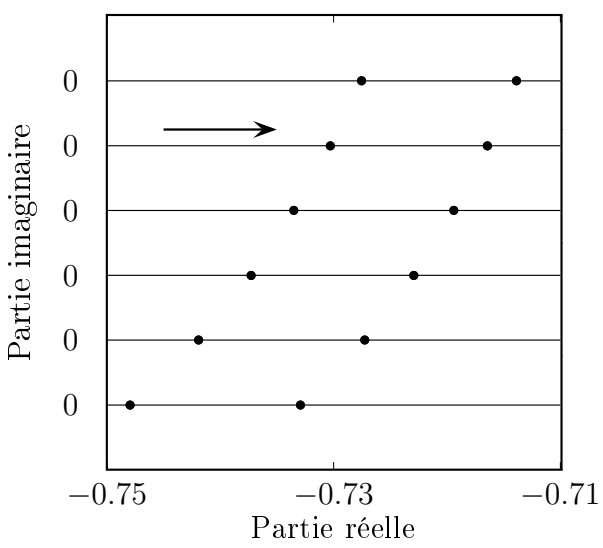


(b) Évolution temporelle de l'erreur de modélisation des réponses impulsionnelles de la fibre auditive.

Figure 2.10 Réponses impulsionnelles du filtre binomial compressif et évolution temporelle de l'erreur de modélisation pour l'Unité 86100u25.



(a) Déplacement des pôles du filtre binomial compressif.



(b) Déplacement des zéros du filtre binomial compressif

Figure 2.11 Déplacement des pôles et zéros dans le plan z du filtre binomial compressif quand le niveau d'excitation change (la direction pointée par la flèche indique un niveau d'excitation croissant).

conséquences : (1) le gain maximal du spectre du filtre cBIT diminue et (2) la fréquence de résonance migre vers les basses fréquences. Les zéros sont purement réels et de multiplicité 2 chacun : ils se déplacent aussi vers zéro où ils suivent les pôles dans leur déplacement pour assurer des FIs dont les trajectoires sont indépendantes du niveau d'excitation.

Le filtre cBIT est stable, causal et à phase minimale. Ce sont des propriétés utiles si ce filtre est utilisé dans un scénario d'analyze-synthèse. Nous étudions dans la section suivante l'aptitude du filtre cBIT à modéliser les RIs du NA du même chat mais pour différentes fibres auditives.

2.3.3 Résultats de modélisation de l'Unité 86100

Trente RIs du NA pour différents niveaux d'excitation et différentes MF du même chat sont utilisés dans cette section pour valider le filtre cBIT. Les RIs du modèle sont générés utilisant l'équation (2.21) en utilisant la même procédure que celle utilisée dans la section 2.2.2. L'ajustement du modèle est effectué dans le domaine temporel et tous les paramètres ont été définis comme étant dépendants du niveau d'excitation, sauf pour le cas du paramètre γ assurant ainsi des FIs dont les trajectoires sont indépendantes du niveau d'excitation.

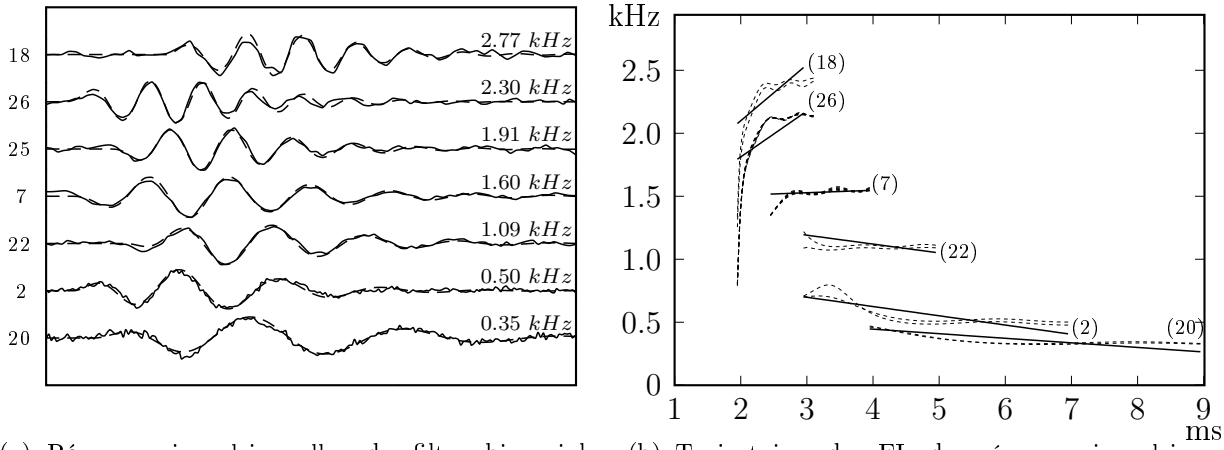
Tableau 2.2 Erreur de modélisation (en dB) des réponses impulsionnelles des fibres de l'Unité86100. Les performances du filtre gammachirp compressif (CGC) et le filtre gammachirp analytique (AGC) sont aussi données.

Nř. fibre	N	MF	e(CBIT)	e(AGC)	e(CGC)
20	2	351	-11.2	-10.0	-9.5
2	2	508	-11.7	-10.7	-10.9
22	4	1094	-11.7	-10.3	-9.8
7	5	1602	-12.1	-12.7	-10.4
25	9	1914	-13.4	-11.0	-12.0
26	2	2305	-11.0	-10.2	-11.2
18	2	2773	-10.1	-8.5	-8.8

Le tableau 2.2 présente les erreurs d'ajustement (exprimées en dB) du filtre cBIT, du filtre AGC ainsi que celles du filtre cGC. La deuxième colonne de ce tableau donne le nombre des RIs par fibre alors que la troisième colonne donne les fréquences de résonances de ces fibres. Les erreurs concernant le filtre AGC ainsi que celles du filtre cGC ont été reportées des tableaux IV et V de [Irino et Patterson, 2001] respectivement.

Le tableau 2.2 montre que le filtre cBIT peut modéliser les RIs de l'Unité 86100 avec une erreur moyenne 1.2 dB plus petite que celle réalisée soit par le filtre AGC ou par le filtre cGC. L'erreur de modélisation de la fibre 25 en utilisant le filtre GT est de -10.1 dB

([Irino et Patterson, 2001]). Pour cette même fibre, le filtre cBIT réalise une erreur de modélisation 2.5 dB plus petite que celle réalisée par le filtre AGC et 3.3 dB que celle obtenue avec le filtre GT.



(a) Réponses impulsionnelles du filtre binomial compressif (en ligne pointillée) pour le niveau d'excitation de 80 dB SLP. Le numéro des fibres est donné sur l'axe des ordonnées.

(b) Trajectoires des FIs des réponses impulsionnelles du filtre binomial compressif (en ligne pointillée). Les droites de régression des FIs des fibres pour le niveau d'excitation de 80 dB SLP est représentée par une ligne continue. Le numéro des fibres est donné entre parenthèses.

Figure 2.12 Réponses impulsionnelles du filtre binomial compressif ainsi que les trajectoires de leurs fréquences instantanées pour l'Unité86100.

La figure 2.12 présente les RIs du cBIT pour 7 fibres auditives de la même unité ainsi que leurs trajectoires [5]. La figure 2.12(b) présente les droites de régression représentant les trajectoires de FIs des RIs du NA pour les niveaux d'excitation de 80 dB SPL : les trajectoires du cBIT passent à travers les bornes délimitées par les droites de régression. Les trajectoires sont également indépendantes du niveau d'excitation. Les fibres 26 et 18 ont des FIs qui s'élèvent rapidement et atteignent rapidement leurs fréquences asymptotiques. La pente de la trajectoire au début de la réponse impulsionnelle est très élevée et décroît ensuite rapidement. Ceci est différent des autres fibres où la variation de la pente de la trajectoire est moins abrupte. Le comportement de ces fibres a également été rapporté dans [Carney *et coll.*, 1999] et [Irino et Patterson, 2001] : ces fibres montrent des trajectoires oscillantes plutôt que des trajectoires lisses.

Dans cette section, on a montré que le filtre cBIT permet de modéliser fidèlement les RIs du NA. Ceci est montré en comparant les propriétés des RIs à celles du NA :

1. La réponse impulsionnelle du cBIT ressemble à celle du NA.

5. Les trajectoires des FIs ont été estimées comme étant les dérivées de la transformée de Hilbert des RIs.

2. Les trajectoires des FIs du cBIT sont indépendantes du niveau d'excitation.
3. La fréquence de résonance du cBIT dévie vers les basses fréquences quand le niveau d'excitation augmente.
4. Le filtre cBIT présente une compression non linéaire au voisinage de la MF.

On s'attarde dans la dernière section de ce chapitre sur la comparaison entre le modèle proposé et les modèles auditifs populaires. On va s'intéresser essentiellement aux propriétés du NA que ces modèles modélisent et à la complexité en terme d'implémentation digitale. Cette comparaison a pour but de démontrer l'avantage du cBIT en terme de réduction de complexité pour des performances comparables.

2.4 Comparaison avec d'autres banc de filtres auditifs

Le filtre BIT est lié au filtre GT, puisque la différence entre les deux réside dans le début de la RI. En effet, le développement en série de Taylor au voisinage de zéro des réponses impulsionnelles de ces derniers filtres donne :

$$\mathrm{GTF}_3(t,\gamma) \approx t^3 - \gamma \times t^4 + O(t^5) \tag{2.22}$$

$$\mathrm{BIT}_n^k(t,\lambda) \approx t^n - \lambda \times (k + 1/2n)t^{n+1} + O(t^{2+n}) \tag{2.23}$$

Quand $n = 3$, choisir une valeur de $\lambda = \gamma/(k + 3/2)$ résulte en un filtre BIT dont le spectre est quasiment identique à celui du GT du 4ième ordre. Un exemple est donné sur la figure 2.13 où le spectre du BIT est délibérément décalé de 1 dB.

La différence entre les deux spectres est aussi donnée : la différence entre les deux filtres est localisée au voisinage de la fréquence de résonance et est inférieure à 0.5 dB. Même si les deux filtres sont très proches, le filtre BIT a l'avantage de posséder une représentation spectrale plus compacte et un degré de liberté de plus pour le même ordre n. Ces deux avantages ont permis au BIT de modéliser de façon plus appropriée les RIs du NA.

Le filtre de cBIT peut être considéré comme un filtre qui appartient à la famille des filtres présentés par [Katsiamis *et coll.*, 2007] : La transformée de Laplace du cBIT pour $n = 2$ est une mise en cascade d'un filtre OZG (d'ordre 1), d'un filtre APG (d'ordre 2) et d'un filtre pôles-zéros en cascade où les pôles et zéros se déplacent par le même taux (PZFC5) (d'ordre 1). Le tableau 2.3 donne une comparaison des familles de filtres auditifs les plus populaires dans la littérature. Cette comparaison est basée sur le critère de simplicité de description (dans le domaine temporel, celui de Laplace et celui de la transformée en z), la ressemblance de la RI du modèle à celle du NA, et le nombre de coefficients de

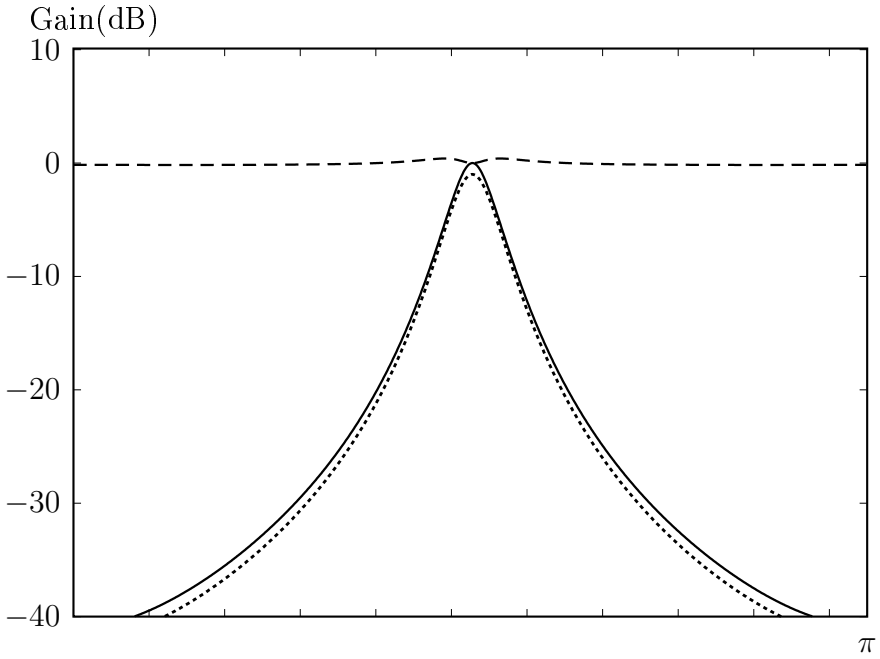


Figure 2.13 Comparaison entre le spectre du filtre Gammatone (pointillé) et celui du filtre binomial (ligne continue) où ce dernier est décalé de 1dB. La différence entre les deux est donnée en ligne discontinue.

Tableau 2.3 Comparaison entre les familles des filtres auditifs.

	BIT	GammaTones			Cascade (3 Filtres/ERB)			Zilaney
Filtres	$cBIT_2$	GT	AGC	cGC	APFC	PZFC	PZFC5	[Zilany, 2006]
Simple	$t/s/z$	t	t	t	s/z	s/z	s/z	s/z
Enveloppe	+	+	+	+	≈	+	+	+
Fréq. inst	+	−	+	+	−	≈	+	+
Compression	+	−	≈	+	+	+	+	+
N. coefficients	15	14	80	80	12	25	25	34

l'implémentation digitale[6]. On a exclu de ce tableau certaines familles comme les filtres DRNL de [Lopez-Poveda et Meddis, 2001; Meddis *et coll.*, 2001] ou le modèle proposé par [Verhulst *et coll.*, 2012] : Ces modèles ont été exclus de cette comparaison puisqu'ils n'ont pas d'implémentation digitale.

Le tableau 2.3 montre que le filtre cBIT permet de modéliser fidèlement les RIs du NA et ce avec un nombre de coefficients réduit. Comparativement au filtre cGC, le cBIT permet une réduction de complexité par un facteur de $3/16$ et une réduction de complexité de l'ordre de $1/2$ comparativement au modèle de [Zilany et Bruce, 2006]. Bien que le filtre cBIT peut être une alternative intéressante aux modèles plus sophistiqués, il faut être conscient de

6. Les symboles +, ≈ et − représentent une validation satisfaisante d'un critère donné, une validation partielle d'un critère donné et l'absence complète d'un critère donné respectivement.

la portée de cette étude. En fait, l'investigation sur les performances du modèle proposé a été limitée à : (1) des niveaux sonores modérés et (2) aux MFs inférieures à 4 kHz[7]. Autres propriétés comme le retard, la suppression dû à la présence d'une deuxième tonalité et la nature de la phase ne sont pas concernés par ce chapitre. En fait, le principal avantage du filtre cBIT montré dans ce chapitre est la capacité de ce dernier à modéliser fidèlement les RIs du NA et ce avec un nombre minimal de coefficients.

2.5 Conclusion

Remplacer la distribution gamma par une loi binomiale dans l'expression du filtre GT donne lieu au filtre binomial. Celui-ci peut être facilement modifié afin d'y inclure une asymétrie spectrale et une compression non-linéaire.

La RI du filtre BIT a été comparée à celles des fibres nerveuses auditives (voir tableau 2.2) : Avec un pôle supplémentaire, le filtre cBIT se transforme en un modèle dont la réponse est fidèle aux RIs du NA en termes d'erreur temporelle, de compression et en terme de trajectoire des FIs. Le modèle, implémenté comme un filtre IIR utilisant seulement 15 coefficients par fibre, est une alternative moins complexe à des modèles plus sophistiqués.

Vu la complexité réduite du filtre cBIT, on propose dans le chapitre suivant, un banc de filtres pour codage audio basé sur ce dernier. Pour ce faire, un banc de filtre adaptatif est conçu pour simuler la réponse fréquentielle du NA humain. La variation des paramètres du banc de filtres est validée en se basant sur des expériences de masquage fréquentiel. Encore une fois, le filtre cBIT se distingue par sa complexité réduite comparativement aux autres banc de filtres auditifs pour des performances comparables.

7. En raison de l'absence de données pour des MFs supérieures à 4kHz.

CHAPITRE 3

Nouveau banc de filtres auditifs dynamiques à base des filtres binomiaux

Les modèles des filtres auditifs publiés dans la littérature comprennent à la fois :

1. ceux motivés par les expériences psychoacoustiques, telles que la détection des tonalités en présence de bruit masquant.
2. ceux qui sont motivés par la reproduction de la réponse mécanique de la membrane basilaire ou celles des réponses impulsionnelles du nerf auditif.

Ceci ne conduit pas nécessairement aux même modèles. Le travail présenté dans ce chapitre consiste à fournir un modèle qui peut non seulement modéliser les réponses mécaniques de la membrane basilaire, synthétiser les réponses impulsionnelles du nerf auditif mais aussi expliquer les expériences du masquage fréquentiel. Étant donné qu'il y a plusieurs étapes de traitement neural entre la cochlée « mécanique » et la perception, il ne serait pas surprenant de trouver que les paramètres diffèrent selon les propriétés cherchées à modéliser. Il semble cependant que la cochlée joue un rôle suffisamment important pour expliquer certaines manifestations de la perception. Dans le chapitre 2, on a présenté le filtre binomial compressif comme étant un modèle à faible complexité des réponses impulsionnelles du nerf auditif. Les performances de ce filtre ont été validées par rapport aux réponses impulsionnelles de la membrane basilaire de chats. Dans ce chapitre, on valide le même modèle par rapport aux expériences de masquage fréquentiel caractérisant ainsi la forme des filtres auditifs humains et fournissant ainsi un filtre auditif à faible complexité permettant d'expliquer certaines manifestations de la perception auditive tout en étant en concordance avec les données physiologiques.

3.1 Architecture du banc de filtres d'analyze et de synthèse

Dans [Duifhuis, 2004], l'auteur résume les modèles cochléaires et les divise en deux catégories : (1) classe des modèles des lignes de transmission et (2) modèles de la classe des filtres. Plus précisément : « La principale différence est que les modèles de la classe 1 prennent en compte le couplage entre les éléments du système, tandis que dans la classe

2 les canaux sont indépendants et le couplage est complètement déterminé par l'entrée commune ». Les filtres en cascades tels que ceux proposés par Katsiamis *et coll.* [2007] (les filtres APFC, OZG et PZFC) fournissent des modèles où le couplage entre les canaux peut se faire facilement puisqu'il suffit d'introduire n'importe quel type de traitement entre les différents canaux. Cependant cette famille de filtres (en cascade) souffre de deux inconvénients majeurs :

1. L'accumulation des erreurs de calcul d'un étage à un autre nécessite une precision numérique élevée. Généralement ces modèles utilisent un nombre de 100 filtres élémentaires (avec 50% de recouvrement) pour couvrir la bande de fréquence d'intérêt.

2. L'ajustement de ces modèles aux données physiologiques ou psychoacoustiques nécessite un nombre élevé de blocs élémentaires (pour modéliser un système continu à savoir une ligne de transmission). Ceci devient un dilemme gênant quand ces modèles sont utilisés avec un nombre de blocs différent de celui utilisé pour leur ajustement.

Pour ces dernières raisons, le modèle proposé est basé sur un banc de filtres cBIT disposés en parallèle. La structure en parallèle proposée est présentée par la figure 3.1. Le signal $s(n)$ à l'entrée du banc de filtres est passé à travers les filtres d'analyze H_i pour donner les signaux $y_i(n)$. Le signal reconstruit $\hat{s}(n)$ est synthétisé à partir des signaux $\hat{y}_i(n)$ en utilisant les filtres de synthèse G_i. Quand les filtres d'analyse sont à minimum de phase, et quand $\hat{y}_i(n)$=$y_i(n)$:

$$G_i = H_i^{-1} \text{ et } \hat{s}(n) = s(n) \tag{3.1}$$

Il est évident que cette structure n'est pas efficace en terme de compression entropique puisque cette dernière crée une redondance proportionnelle au nombre de filtres que contient le banc de filtres. Dans ce chapitre on va se consacrer plutôt à démontrer que le filtre cBIT permet de caractériser aussi le système auditif humain. Les paramètres du banc de filtres à base de cBIT seront ajustés aux expériences de masquage fréquentiel pour caractériser ce filtre et démontrer qu'il est capable d'expliquer le phénomène du masquage tout en étant consistant avec les observations biologiques. On discutera l'utilisation de ce même banc filtres pour la compression audio dans les chapitres suivants.

On a montré dans le chapitre 2, que le filtre auditif est dynamic et compressif : la forme du filtre et ses paramètres changent en fonction de la nature et l'intensité du signal à son entrée. Deux approches peuvent être suivies pour modéliser ce couplage entre l'entrée du filtre et sa sortie. La figure 3.2 présente les deux architectures possibles (en boucle ouverte/fermée).

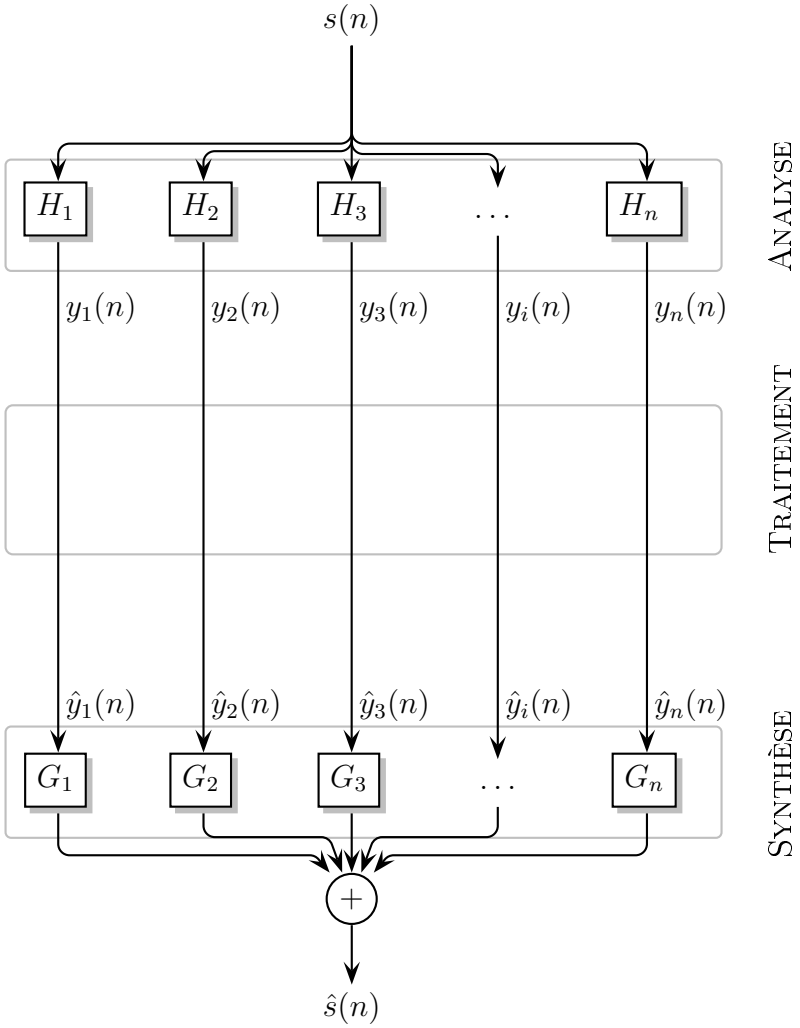
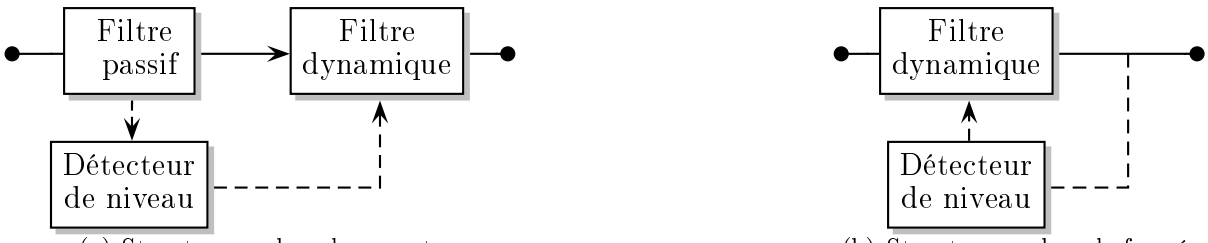


Figure 3.1 Structure en parallèle du filtre auditif proposé.



(a) Structure en boucle ouverte.

(b) Structure en boucle fermée.

Figure 3.2 Adaptation des paramètres du banc de filtres au signal d'entrée. La ligne continue représente le signal d'entrée alors que celle en pointillée représente les paramètres de contrôle.

La figure 3.2(a) présente la structure en boucle ouverte. Pour cette structure, les paramètres du filtre dynamique sont adaptés et ce en se basant sur le niveau du signal à la sortie du filtre passif qui lui y est associé. La structure en boucle ouverte est généralement

plus facile à implémenter mais puisque le signal doit être filtré par deux filtres (passif et actif), la complexité est accrue.

Quant à la structure en boucle fermée présentée par la figure 3.2(a), c'est le niveau de puissance à la sortie du banc de filtres actifs qui détermine la forme de ces derniers. La complexité d'implémentation est réduite dans ce cas puisque le signal d'entrée n'est filtré qu'une seule fois.

Dans les sections suivantes, on va valider le banc de filtres agencé en boucle fermée puisque celle ci représente une architecture ayant une complexité moins élevée que celle pour la structure en boucle ouverte.

3.2 Dérivation des paramètres du banc de filtres

3.2.1 Le modèle du masquage fréquentiel

Une des nombreuses expériences permettant de caractériser le canal auditif humain est celle où un nombre de personnes essaient de détecter une tonalité en présence d'un bruit masquant à bande étroite. Un schéma de l'expérience est donné par la figure 3.3.

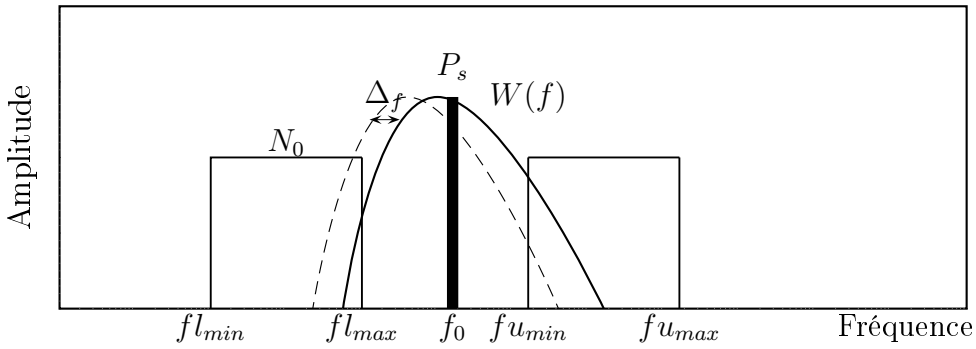


Figure 3.3 Détection de tonalité en présence de bruit masquant.

Dans cette expérience, l'intensité P_s de la tonalité est variée pour déterminer la puissance N_0 du bruit nécessaire pour que cette première soit juste audible et ce pour différentes largeurs de bruit masquant $[fl_{max}, fu_{min}]$. Cette méthode a été utilisée pour caractériser le filtre auditif chez les humains [Webster *et coll.*, 1952]. La même méthode devint plus importante avec les travaux de [Patterson, 1986; Patterson et Nimmo-Smith, 1980] où ils ont démontré que les auditeurs utilisent un filtre auditif qui maximise le rapport signal sur bruit et non un filtre centré autour de la fréquence centrale f_0 de la tonalité. Le masquage d'une tonalité par un bruit blanc à bandes étroites est préféré aux autres types de masquage (par example le masquage d'une tonalité par une autre) quand il s'agit de la

caractérisation de la forme du filtre auditif et ce pour plusieurs raisons. Premièrement, l'expression de la valeur de la tonalité peut être exprimée de façon simple en fonction du niveau du bruit (voir l'équation (3.2)). Deuxièmement, les différentes largeurs du bruit masquant permettent de caractériser indépendamment la forme du filtre auditif vers les basses fréquences de sa forme vers les hautes fréquences. Troisièmement, ce type de masquage permet de prendre en considération la déviation de la fréquence centrale du filtre auditif utilisé par les auditeurs (dans le but de maximiser le rapport signal sur bruit).

Plusieurs auteurs ont répété et étendu ces mêmes expériences pour couvrir plusieurs sujets, différentes fréquences et différentes formes de bruit [Baker *et coll.*, 1998; Glasberg et Moore, 1990; Lutfi et Patterson, 1984; Rosen et Baker, 1994]. D'autres ont donné des analyses de plus en plus sophistiquées pour définir des formes de filtres auditifs qui permettent de prédire les données expérimentales (à savoir la valeur de P_s)[Glasberg et Moore, 2000; Irino et Patterson, 2006b, 1996, 2001]. Leurs données et leurs méthodes sont utilisées dans ce chapitre pour caractériser le banc de filtres proposé.

Deux ensembles de données, couvrant un large éventail de niveaux de bruit et de fréquence, avec plusieurs sujets dans chaque ensemble, ont été utilisés pour ajuster et comparer les différents modèles de filtres auditifs cités dans ce chapitre. Le premier ensemble contient des données collectées pour neufs sujets et pour neuf différentes fréquences en utilisant du bruit blanc [Baker *et coll.*, 1998]. Le deuxième ensemble contient des données collectées par Glasberg et Moore [2000] en utilisant un bruit uniformément distribué (un bruit qui produit le même niveau d'excitation par ERB). Pour la plupart des travaux cités ci-haut, y compris pour le présent travail, l'ajustement des filtres a été réalisé en utilisant les seuils moyens à travers les sujets de chaque ensemble de données totalisant ainsi 1277 expériences différentes. Les fréquences utilisées durant ses expériences couvrent la bande [250 Hz, 6 kHz].

Si la pente du spectre du bruit masquant à la fréquence de coupure est très raide, il est possible d'écrire une fonction qui corrèle le seuil de masquage de la tonalité à la forme du filtre auditif. Cette relation constitue la base de la procédure d'ajustement utilisée pour calculer la forme du filtre auditif [Glasberg et Moore, 1990; Lyon, 2011b; Moore *et coll.*, 1990; Patterson, 1976]. Si la forme du filtre auditif peut être représentée par une fonction de pondération $W(f)$, alors le seuil de masquage de la tonalité peut être prédit par :

$$\hat{P}_s = K \times \int_{\Delta f} N_0 \times W(f) df \tag{3.2}$$

où pour des raisons de clarté :

$$\begin{cases} P_s = P_s(f_0, N_0, fl, fu) & W(f) = W(f)_{f_0} \\ K = K(f_0) & N_0 = N_0(f_0, fl, fu) \\ \Delta_f = [fl_{\min}\ fl_{\max}] \cup [fu_{\min}\ fu_{\max}] & \end{cases}$$

avec K un paramètre qui dépend uniquement de la fréquence et $\hat{P}_s$ le niveau de la tonalité prédit par le modèle. Ce modèle est appelé « modèle de puissance du masquage » parce qu'il suppose un bruit stationnaire. La caractérisation du modèle proposé dans ce chapitre se base sur l'équation 3.2 pour ajuster les différents paramètres du modèle.

3.2.2 Algorithme d'ajustement du banc de filtres

Dans cette section, l'*ajustement* du modèle fait référence au processus utilisé pour trouver les « meilleures » valeurs des paramètres du modèle proposé de telle sorte que si l'équation (3.2) est utilisée pour déterminer la valeur de $\hat{P}_s$, l'erreur quadratique moyenne entre $\hat{P}_s$ et P_s est la plus petite possible. Cette recherche constitue un problème de minimisation quadratique mais puisque le système (lien entre $\hat{P}_s$ et $W(f)$) est non-linéaire, la recherche est plus compliquée.

Pour l'optimisation non-linéaire, la méthodologie de [Glasberg et Moore, 2000; Irino et Patterson, 2006b, 2001] est suivie avec plusieurs modifications en utilisant l'algorithme de Levenberg-Marquardt [Moré, 1978] appliquées aux données de [Baker *et coll.*, 1998; Glasberg et Moore, 2000] : le travail présenté dans cette thèse n'aurait pas été possible sans l'aide généreuse de tous ces auteurs.

Chaque banc de filtres a ses propres paramètres à ajuster en plus de trois paramètres à savoir :

1. La fréquence centrale du filtre w_0 qui doit être déterminée pour maximiser le RSB à la sortie du filtre.
2. Le niveau du bruit de fond P_0 : un paramètre qui représente le niveau du bruit interne au niveau de l'oreille. Ce bruit est ajouté à l'entrée du filtre auditif avant de calculer P_s.
3. La constante de détection K.

L'oreille externe et moyenne affectent le spectre du bruit masquant et donc les niveaux des bandes de bruit à la sortie du filtre auditif, en particulier lorsque la fréquence de la tonalité est faible ou élevée. [Glasberg et Moore, 1990; Rosen et Baker, 1994] ont montré

que l'inclusion d'une fonction de transfert qui représente le filtrage pré-cochléaire réduit l'erreur quadratique.

L'algorithme d'estimation des tonalités en utilisant un modèle M est présenté par l'algorithme 3.1.

Algorithme 3.1 : Algorithme d'estimation des tonalités $\hat{P}_s$.

Entrées : Données des expériences psychoacoustiques, M : Modèle du banc de filtres
Sorties : $\xi(M)$: erreur de prédiction des tonalités.
pour *chaque expérience* **faire**
  Générer H_i, les spectres des filtres auditifs en utilisant le modèle M.
  Calculer $W_i(f)$ les spectres affectés par la fonction de transfert de l'oreille externe et moyenne.
  Inclure la fonction de transfert du bruit de fond P_0(f).
  Trouver le filtre $W(f)$ parmi ces filtres qui maximise le rapport signal sur bruit.
  Estimer $\hat{P}_s$ en utilisant l'équation (3.2).
  $e(M) = P_s - \hat{P}_s$.
fin
Trouver la valeur de K pour minimiser $\xi(M) = e(M) - K$ et ce pour toute les expériences concernées ;

L'erreur $\xi(M)$ est ensuite utilisée par l'algorithme de Levenberg-Marquardt comme étant la fonction d'erreur non-linéaire à minimiser. Il faut noter que la constante K ne dépend que de la fréquence et n'intervient nullement dans la détermination de la forme du filtre auditif.

3.2.3 Modélisation du filtre auditif chez les humains

Le système auditif peut être modélisé par un ensemble infini de filtres auditifs dynamiques qui couvrent le spectre auditif. Comme illustré au chapitre 2, certaines propriétés de la RI du NA dépendent de la fréquence tel que la fréquence centrale, la largeur du filtre auditif, la trajectoire des FIs. D'autres paramètres dépendent du niveau de l'excitation par exemple la déviation de la fréquence centrale du filtre où la compression au niveau du gain maximal.

Bien que les non-linéarités se manifestent de diverses manières dans le système auditif, il est toujours possible de modéliser une bonne partie de ces non-linéarités avec des modèles quasi-linéaires : modèles qui peuvent être décrits comme étant des filtres linéaires mais avec des paramètres qui dépendent de façon non-linéaire du niveau de l'excitation. De tels modèles peuvent capter les principaux effets de masquage et la compression entre entrée et sortie associée aux mécanismes cochléaires. On décrit dans les parties suivantes la dépendance de la formes des filtres proposées par rapport à leurs fréquences de résonance ainsi que par rapport au niveau de l'excitation.

La transformée de Laplace du filtre binomial compressif d'ordre 2 (cBIT_2) est donnée par :

$$H_{\mathrm{cBITF}_2}(s) = \frac{K(s+z_1)[(s+z_1)^2 - (a^2+3w_0^2)]}{\prod_{k=0}^{3}[(s+p_k)^2+w_0^2]} \tag{3.3}$$

Avec $p_k = b + k \times b_{rat}$, $z_1 = a + b$ et w_0 la fréquence de résonance du cBIT_2. On décrit dans les sections suivantes la dépendance des pôles et zéros du cBIT_2 par rapport à la fréquence du filtre auditif et par rapport au niveau de l'excitation à son entrée. On a choisit le filtre cBIT_2 comme example, mais la même notation ainsi que le raisonnement restent valident pour n'importe quel ordre.

Dépendance par rapport à la fréquence : Quand le niveau d'excitation est faible, le filtre auditif est symétrique et la valeur de sa largeur de bande équivalente centrée autour de f_c est approximée par [Glasberg et Moore, 1990] :

$$\mathrm{ERB}(f_c) = 24.7 + 0.1079 \times f_c \tag{3.4}$$

$$\mathrm{ERB}_{\mathrm{rate}}(f_c) = 24.1 \times \log_{10}(1 + 0.00437 \times f_c) \tag{3.5}$$

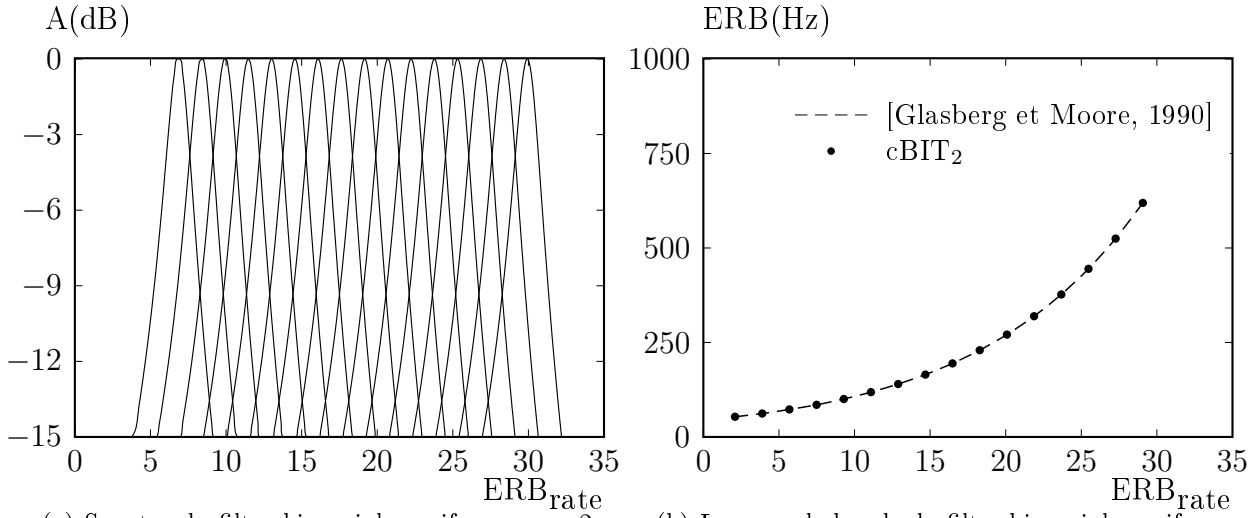
Dans le cas du GT d'ordre 4 (tel que défini par l'équation (2.5)), le lien entre le facteur d'amortissement b_{gt} et la largeur de bande équivalente est donné par [Patterson, 1976] :

$$b_{\mathrm{gt}} = 1.019 \times \mathrm{ERB}_{\mathrm{rate}} \tag{3.6}$$

Comme démontré à la section 2.4, le cBIT_2 peut donner un filtre dont le spectre est égal à celui du GT. On peut alors définir le filtre cBIT_2 comme étant un filtre GT quand le niveau d'excitation est faible (Pour les niveaux d'excitation faibles le filtre auditif est symétrique). Pour ce faire, il suffit de choisir $b = -a = b_{\mathrm{gt}}$ et $b_{rat} = 0$. Avec ces valeurs, le cBIT_2 est un GT tel qu'illustré par la figure 3.4. De façon plus générale, pour un filtre binomial d'ordre n, *Binonial-tone filter* (BIT_n) la valeur de b_0 donnant la largeur de bande équivalente est donnée par :

$$b_0 = v(n) \times \mathrm{ERB}_{\mathrm{rate}} \tag{3.7}$$

Avec $v = \begin{pmatrix} 0.638 & 0.849 & 1.019 & 1.164 & 1.293 & 1.329 \end{pmatrix}$. Pour un filtre binomial compressif d'ordre n (cBIT_n), il suffit d'utiliser $v(n+1)$.



(a) Spectre du filtre binomial passif pour $n = 2$.

(b) Largeur de bande du filtre binomial passif pour $n = 2$.

Figure 3.4 Spectres et largeurs de bande équivalente du banc de filtres proposé quand le niveau d'excitation est faible.

Sur la figure 3.4(a), le spectre d'un banc de 16 filtres cBIT_2 passifs est donné. Les filtres cBIT_2 passifs tels que décrits précédemment sont symétriques et leur ERBs sont égales à celles du système auditif humain tel qu'illustré par la figure 3.4(b).

Dépendance par rapport au niveau d'excitation : Comme démontré dans le chapitre 2, la RI du NA dépend de l'intensité de l'excitation. Cette excitation a pour effet de modifier la MF, le gain maximal de la RI du NA sans pour autant modifier la trajectoire de ses FIs. On a démontré dans le chapitre précédent que ces contraintes peuvent être satisfaites si le paramètre a est indépendant du niveau d'excitation alors que w_0 et b le sont (équation (3.3)).

Rosen et Baker [1994] donnent une fonction qui approxime la compression du système auditif humain :

$$P_{out} = C_1 + 0.9 \times P_{in} + C_2 \times \left(1 - \frac{1}{1 + \exp(-0.05 * (P_{in} - 50))}\right) \tag{3.8}$$

où C_1 et C_2 sont deux constantes et P_{out} et P_{in} sont la puissance à la sortie du filtre auditif et à son entrée respectivement.

On s'inspire de cette équation pour modéliser la variation de la forme du filtre cBIT (b : facteur d'amortissement de la RI du filtre) ainsi que sa fréquence de résonance (w_0) par rapport à la puissance du signal d'entrée P_{in} :

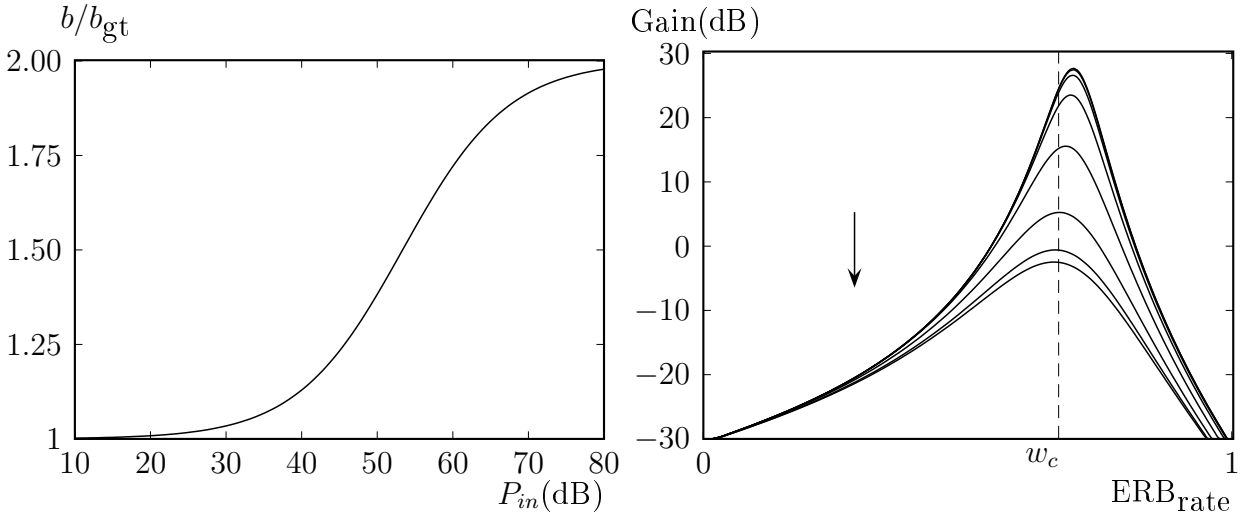
$$b = b_0 \times \left(1 + \frac{C}{1 + \exp(-\delta_P \times (P_{in} - 50))}\right) \tag{3.9}$$

$$w_0 = w_c \times (1 + \delta_w \times (2 - b_{\mathrm{gt}}/b)) \tag{3.10}$$

où C, δ_P et δ_w sont des constantes. L'équation (3.9) est préférée à une approximation par polynôme pour plusieurs raisons. En effet cette première :

- nécessite un nombre de paramètres plus petits que le nombre de coefficients nécessaires à un polynôme pour reproduire la même forme que celle sur la figure 3.5.
- est differentiable par rapport à tous ses paramètres ce qui est une propriété souhaitée pour l'algorithme d'apprentissage 3.1.
- est bornée entre $[b_{\mathrm{gt}}, C \times b_{\mathrm{gt}}]$ est donc appropriée pour compresser une entrée dont la dynamique est élevée assurant ainsi la stabilité du filtre peu importe la valeur de la puissance P_{in}.

La figure 3.5 donne une représentation de cette relation pour les valeurs suggérées par [Rosen et Baker, 1994]. Quand le niveau d'excitation est faible, le cBIT_2 se comporte comme un filtre GT avec une ERB équivalente à celle du système auditif humain. Le gain du filtre cBIT_2 dans ce cas est constant (d'où le terme passif). Quand le niveau d'excitation augmente, le gain du filtre cBIT_2 diminue et sa largeur de bande augmente. Une fois le niveau devenu élevé (supérieur à 90dB), le gain du cBIT_2 ainsi que sa largeur de bande



(a) Modélisation de la compression du système auditif humain.

(b) Spectres du filtre cBITF_2 dynamique.

Figure 3.5 Spectres et gains du filtre cBITF_2 dynamique pour les valeurs suggérées par [Rosen et Baker, 1994] et $\delta_w = -0.01$. La direction pointée par la flèche indique des niveaux d'excitation croissants allant de 10dB à 80dB avec un pas de 10dB.

restent constants. Ce raisonnement reste valide aussi pour tous les filtres cBIT_n et BIT_n. Pour le cas illustré sur la figure 3.5(b), le filtre cBIT_2 réalise une compression moyenne de 0.58 dB/dB. Les valeurs des paramètres C, δ_w, δ_P qui permettent de prédire la valeur de P_s sont à déterminer et ce à partir des expériences de masquage fréquentiels.

Le modèle du banc de filtres est défini par :

$$\begin{pmatrix} a \\ b_{rat} \\ C \\ \delta_P \\ \delta_w \end{pmatrix} = \mathbf{M} \times f_{dep} \tag{3.11}$$

où $\mathbf{M} \in \mathbb{R}_{5,2}$ et $f_{dep} = [1 \ f_x] \in \mathbb{R}_{2,1}$ avec $f_x = \text{ERB}_{\text{rate}}(f)/\text{ERB}_{\text{rate}}(1\text{kHz})$. On rappelle que :

L'équation (3.11) décrit la variation des paramètres du banc de filtres proposé par rapport à la fréquence. L'algorithme de Levenberg-Marquardt est utilisé conjointement avec l'algorithme 3.1 pour trouver la valeur de $\mathbf{M}$ qui minimise $\xi(M)$. On présente dans les sections suivantes les résultats d'ajustement des filtres proposés aux expériences de masquage.

a : Contrôle la position des zéros dans l'equation (2.21) définissant l'allure du filtre vers les basses fréquences.
b_{rat} : Contrôle la cascade des poles dans l'équation (2.21).
C : Contrôle la largeur de bande maximale du filtre.
δ_P : Contrôle la région de compression du filtre.
δ_w : Contrôle la déviation de la fréquence centrale du filtre par rapport à la puissance.

3.3 Résultats expérimentaux

Dans cette section, on compare les performances de deux familles de filtres : les filtre binomiaux (BIT_n) et les filtres binomiaux compressifs (cBIT_n) où n indique l'ordre utilisé pour chaque filtre. On suit la même méthodologie utilisée dans le chapitre 2 et on commence par étudier l'impact du nombre de coefficients du modèle $\mathbf{M}$ sur l'erreur d'ajustement du banc de filtres aux expériences de masquage.

3.3.1 Compromis complexité et erreur d'apprentissage

La variation de la forme du banc de filtres proposé par rapport à la fréquence et ainsi que celle par rapport aux niveaux d'excitation est donnée par les équations (3.9) et (3.11). Le nombre maximal des coefficients de ce modèle est de 10. Pour chaque famille, et pour chaque nombre donné de coefficients, le modèle donnant la plus petite erreur est donné sur la figure 3.6. Puisque, tel que détaillé à la section 3.1, la structure en boucle fermée est plus économique en terme de complexité, cette section présente les résultats quand les filtres binomiaux sont agencés selon la structure présentée par la figure 3.2(b).

La figure 3.6 réaffirme les conclusions déduites au chapitre 2 :

- Le filtre binomial d'ordre 2 (BIT_2) et le filtre binomial compressif d'ordre 3 (cBIT_3) sont les moins performants. Ceci est dû soit au manque de compression au voisinage de la fréquence de résonance soit à leurs formes symétriques. Ces défauts ne peuvent être corrigés en augmentant le nombre de paramètres et fait de ces filtres des modèles relativement inappropriés pour modéliser le système auditif humain.
- Le filtre binomial d'ordre 3 (BIT_3) et le filtre cBIT_2 réalisent les meilleurs scores cependant le filtre cBIT_2 nécessite moins de coefficients que le filtre BIT_3 quand il s'agit d'implémentation digitale. Cette conclusion a déjà été confirmée au chapitre 2 quand les RIs des deux filtres ont été ajustées aux RIs du NA de chats.

Sur la figure 3.7 les performances d'autres familles de filtres sont aussi données. Leurs valeurs ont été rapportées depuis la figure 6 de [Lyon, 2011b].

Le modèle qui performe le moins bien est le APFC, ceci est essentiellement dû à l'absence de zéros dans sa fonction de transfert d'où un manque de contrôle des basses fréquences.

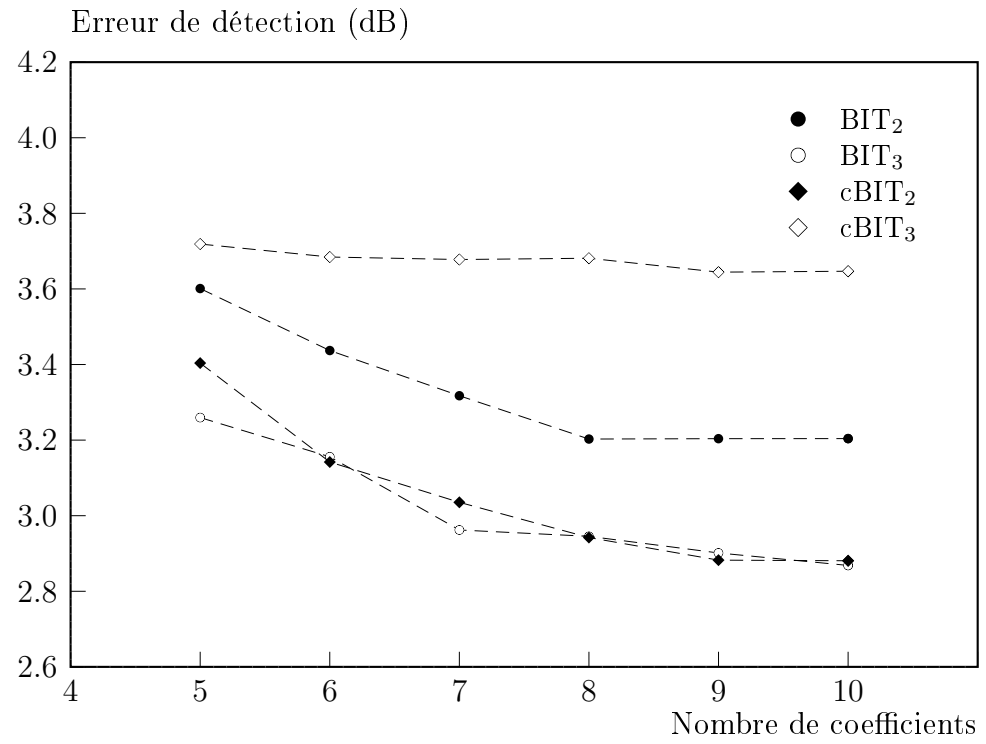


Figure 3.6 Erreur de détection des tonalités pour différentes familles de banc de filtres proposés.

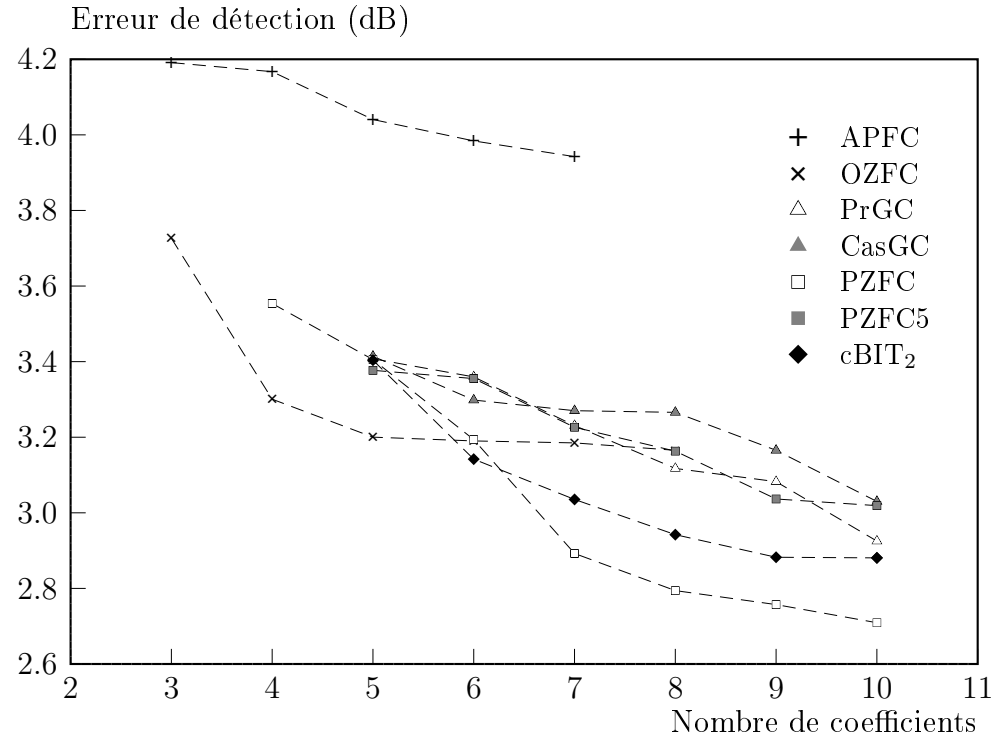
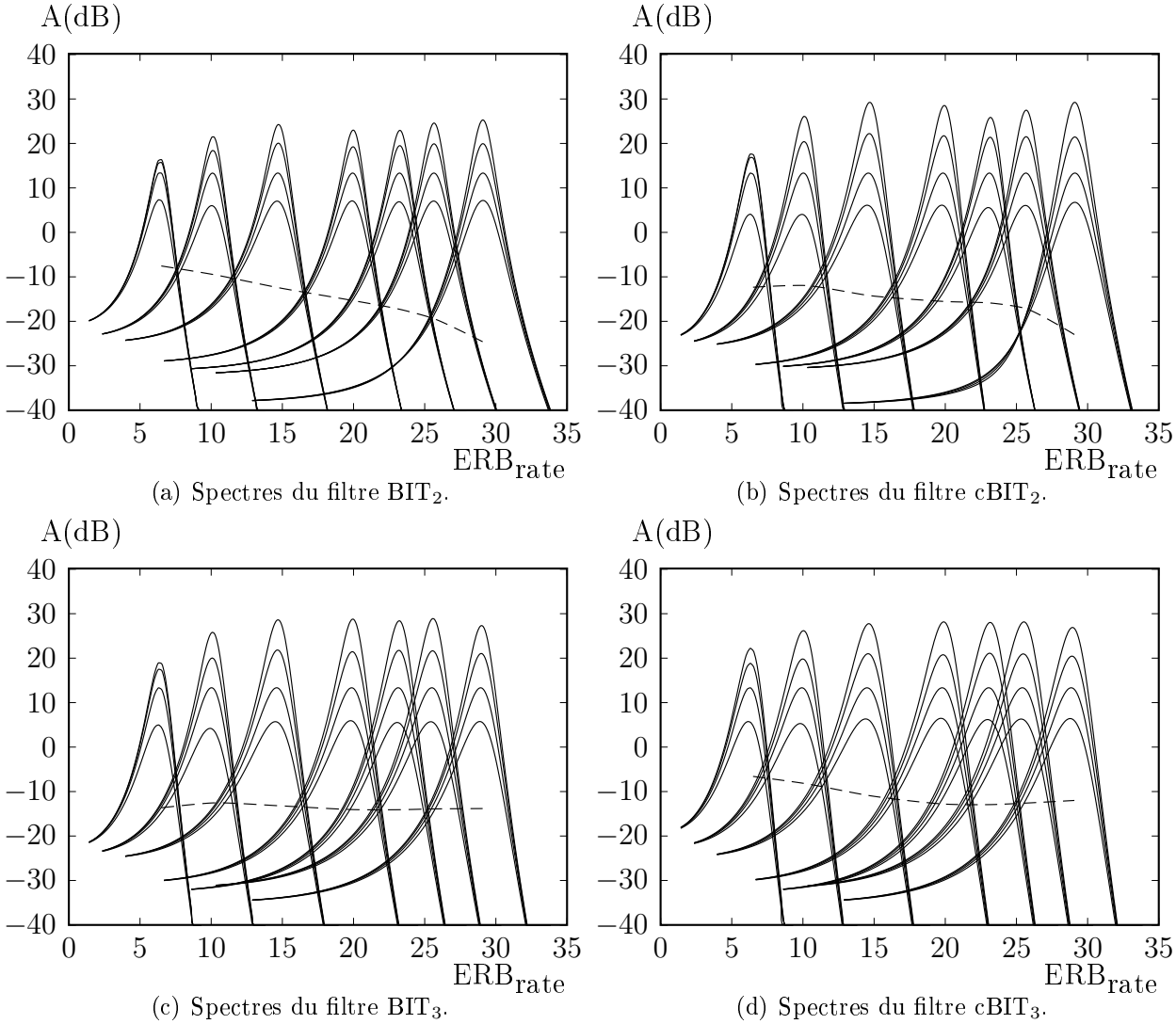


Figure 3.7 Erreur de détection des tonalités par différents modèles. Les données concernant les filtres auditifs autres que le filtre $cBIT_2$ ont été rapportées depuis la figure 6 de [Lyon, 2011b].



(a) Spectres du filtre BIT_2.

(b) Spectres du filtre cBIT_2.

(c) Spectres du filtre BIT_3.

(d) Spectres du filtre cBIT_3.

Figure 3.8 Familles de filtres binomiaux (pour des niveaux d'excitation différents allant de 30 à 70 dBSPL) ajustées aux expériences de masquage. La ligne discontinue représente la valeur du paramètre P_0.

Le filtre tout-pôle en cascade avec un zéro, *one-zero filter cascade* (OZFC) qui est un filtre APFC avec un zéro additionnel dans sa fonction de transfert, réalise un meilleur score. Ce dernier filtre, même s'il réalise une erreur de prédiction relativement réduite, ne peut être considéré comme étant un filtre auditif puisque sa réponse impulsionnelle ne ressemble pas à celle du nerf auditif.

Les filtres cGC à savoir en structure parallèle (PrlGC) ou en cascade (GasGC) se comportent relativement de la même façon par rapport au nombre de coefficients et réalisent sensiblement la même erreur de prédiction. Cependant, ces filtres nécessitent un nombre élevé de coefficients pour une implementation digitale ce qui les rend moins appropriés pour la discipline du codage audio (tableau 2.3).

Les filtres en cascade PZFC5 proposés par [Lyon, 2011b] se comportent relativement de la même façon que les autres familles de filtres et réalisent une réduction de complexité par rapport aux cGCs. Le filtre qui réalise la plus petite erreur est le PZFC mais malheureusement il ne peut être considéré comme un filtre auditif puisqu'il viole la contrainte des FIs dont les trajectoires doivent être indépendantes du niveau d'excitation (les pôles et zéros de ce banc de filtre se déplacent indépendamment quand le niveau d'excitation change, ce qui fait que la trajectoire des fréquences instantanées de ces filtres n'est pas constante). Avec ces résultats, le tableau 2.3 peut être complété. Tel que illustré à la section 2, le filtre

Tableau 3.1 Comparaison entre les familles des filtres auditifs.

	BIT	GammaTones			Cascade (3 Filtres/ERB)			Zilaney
Filtres	$cBIT_2$	GT	AGC	cGC	APFC	PZFC	PZFC5	[Zilany, 2006]
Simple	$t/s/z$	t	t	t	s/z	s/z	s/z	s/z
Enveloppe	+	+	+	+	$\approx$	+	+	+
Fréq. inst	+	−	+	+	−	$\approx$	+	+
Compression	+	−	$\approx$	+	+	+	+	+
$e(dB)$ éq. (3.1)	2.9	5.9	3.5	3.0	3.9	2.7	3.0	$\varnothing$
N. coefficients	15	14	80	80	12	25	25	34

$cBIT_2$ représente un compromis raisonnable entre complexité et erreur d'ajustement aux expériences psychoacoustiques. Dans la suite de la thèse, c'est ce modèle particulier qui sera utilisé comme étant le banc de filtre d'analyse/synthèse pour le codage audio.

3.3.2 Choix de modèle : compromis entre performance et surapprentissage

Jusqu'à présent on a présenté les filtres cBITs comme étant des modèles qui permettent de prédire la valeur des tonalités tel que décrit dans la section 3.2.1. Dans la section précédente

on s'est intéressé à la determination des valeurs numériques de la relation, exprimée par l'équation (3.11), qui décrit la forme du filtre qui minimise l'erreur de prédiction. Les données au nombre de 1277 expériences distinctes ont été utilisées pour trouver les valeurs numériques des paramètres de l'équation (3.11). Le nombre de paramètres maximal pour chaque famille de filtres étant de 10 ($\ll 1277$), il semble être justifié de ne pas soupçonner un surapprentissage. Ceci n'est pas moins valide. On effet, le surapprentissage peut se produire même si le nombre de paramètres est limité (Il suffit par example de penser à une régression linéaire avec plusieurs valeurs aberrantes.). Une des méthodes pour vérifier la validité d'un modèle ($f : x \rightarrow \hat{y}$) consiste à calculer le coefficient de détermination donné par :

$$R^2(f) = 1 - \frac{\sigma^2_{\hat{y}-y}}{\sigma^2_y} \tag{3.12}$$

Plus R est proche de 1, plus le modèle décrit fidèlement les données. Dans le cas du cBIT_2 cette valeur varie entre 95% et 97% (À titre de comparaison, pour le filtre GT ce coefficient vaut 86%). Une autre méthode appelée validation croisée consiste à subdivider les données en deux groupes : le premier utilisé pour entraîner le modèle, l'autre pour valider le pouvoir de généralisation de ce dernier. La figure 3.9, présente les résultats pour la famille de filtres présentées dans ce chapitre. Cette figure expose le dilemme existant

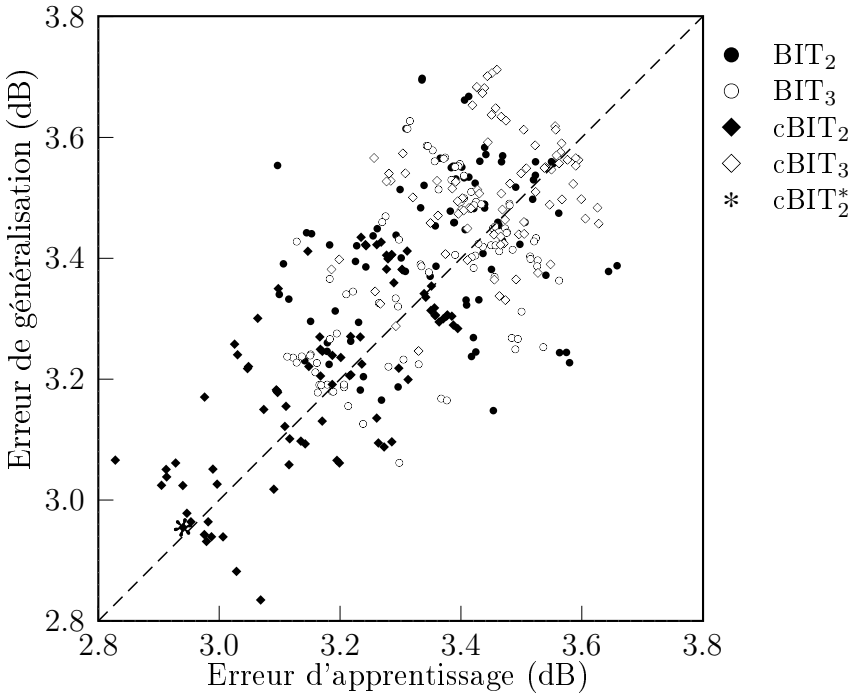
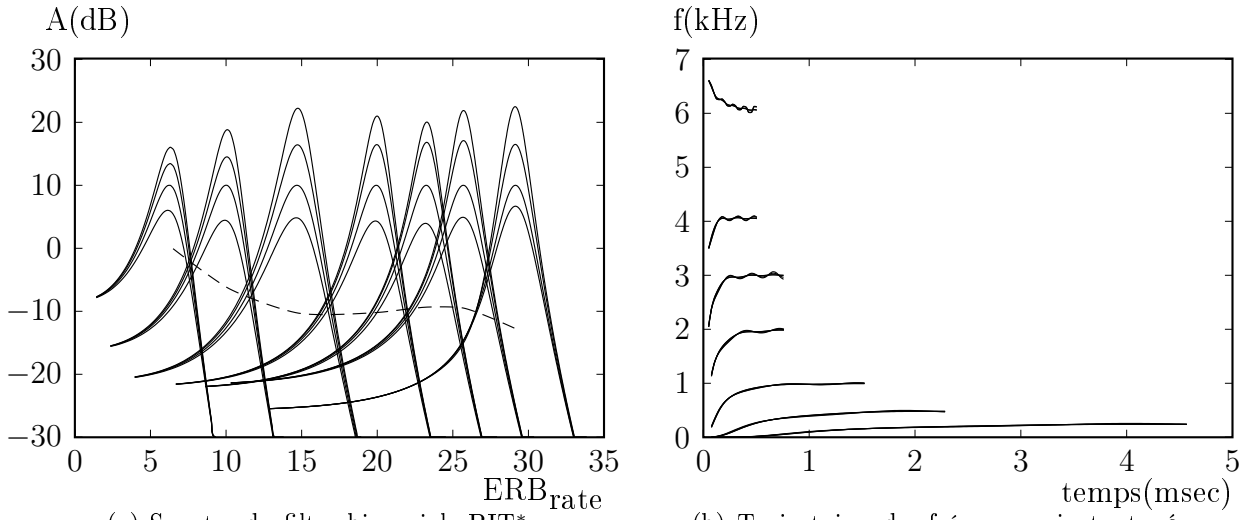


Figure 3.9 Erreur de généralisation vs l'erreur d'apprentissage pour différentes familles de banc de filtres proposées.

entre erreur de prédiction et pouvoir de généralisation : le modèle qui donne l'erreur d'apprentissage la plus petite est celui qui généralise le moins bien.
Le modèle qui semble présenter un compromis entre ces deux critères est celui qui se trouve le plus proche de l'origine du graphe. Le filtre binomial compressif d'ordre 2 réalisant le compromis entre erreur d'apprentissage et de généralisation ($cBIT_2^*$) est représenté par une étoile sur la figure 3.9.

3.3.3 Compression et fréquences instantanées du modèle choisi

Cette section consiste à vérifier que les paramètres du filtres $cBIT_2^*$ ne violent pas les contraintes que doit satisfaire un filtre auditif à savoir compression, fréquences de résonance dépendantes du niveau d'excitation et des trajectoires de fréquences instantanées indépendantes du niveau d'excitation. La figure 3.10(a) représente le spectre et les trajectoires du filtre $cBIT_2^*$ pour différents niveaux d'excitation. Tel qu'illustré sur la figure, les trajectoires des FIs [1] sont indépendantes du niveaux d'excitation. Les réponses impulsion-



(a) Spectre du filtre binomial $cBIT_2^*$.

(b) Trajectoires des fréquences instantanées.

Figure 3.10 Spectres et trajectoires des fréquences instantanées du filtre $cBIT_2^*$ pour des niveaux d'excitation allant de 30 à 70 dB.

nelles du $cBIT_2^*$ quant à elles ressemblent à une distribution Gamma tel qu'illustré sur la figure 3.11. Les temps de passage par zéro étant indépendants des niveaux d'excitation confirment le fait que les trajectoires des RIs sont constantes.

1. Les FIs ont été estimées comme étant les dérivées de la phase de la transformée de Hilbert des RIs.

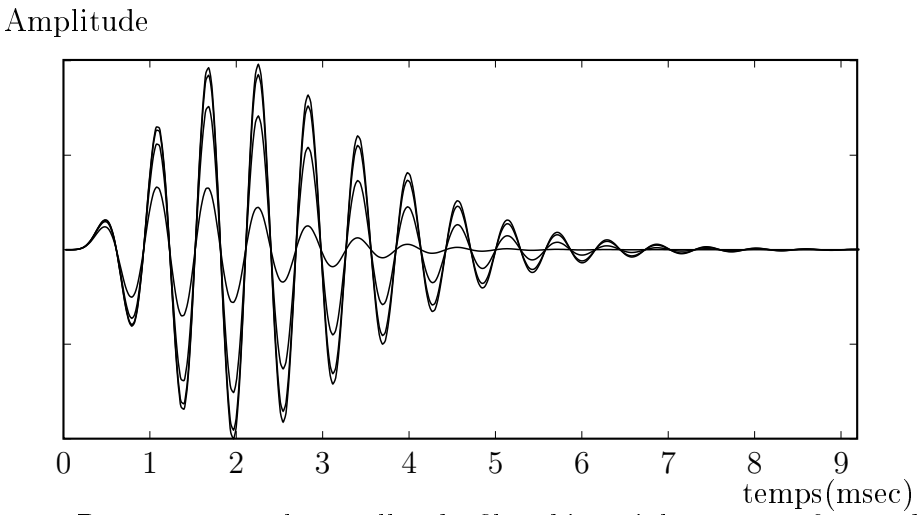


Figure 3.11 Réponses impulsionnelles du filtre binomial compressif pour différents niveaux d'excitation.

3.3.4 Conclusion

Les paramètres des filtres cBITs ont étés ajustés aux experiences du masquage fréquentiel caractérisant le system auditif humain. Les résultats montrent que ces filtres représentent un bon compromis entre complexité d'implémentation et erreur de prédiction des seuils de masquage. Dans ce chapitre, on a décrit le banc de filtres (en boucle fermée) et on a donné l'expression des paramètres qui relient sa forme aux niveaux d'excitation à son entrée : Le cBIT pour une moindre complexité réalise une erreur de prédiction plus petite que les autres familles de filtres auditifs cités dans la littératures (tableau 2.3). Avec seulement 15 coefficients par bande, le cBIT_2^* permet de produire les résultats des expériences du masquage auditif tout en respectant les contraintes physiologiques à savoir la compression, des RIs ressemblant à une distribution Gamma et des trajectoires de FIs indépendantes des niveaux d'excitation. Ceci dit, ce banc de filtres ne peut être utilisé tel que décrit dans ce chapitre pour l'encodage audio puisque, les filtres de ce dernier sont dynamiques et sont positionnés avec un recouvrement de 50% et donc la sortie de ce banc de filtre est très redondante. Les chapitres suivants ont pour but de présenter les méthodes utilisées pour remédier à ce problème.

CHAPITRE 4

Synthèse par inversion des motifs d'excitations auditives

Dans le chapitre 3, on a présenté les cBIT$_2^*$s comme étant des modèles mimant le fonctionnement de la cochlée. On a montré que ces derniers, à moindre complexité, peuvent prédire les seuils de masquage lors des experiences du masquage de tonalités par du bruit blanc à bandes étroites. Ceci dit, le mécanisme d'audition est bien plus complexe et d'autres éléments interviennent lors du processus de perception. Dans le but de se rapprocher encore plus du domaine perceptuel, le banc de filtres proposé est étendu en y incluant des modèles simulant le fonctionnement des cellules ciliées et les neurones qui y sont attachés. Ces derniers peuvent être approximés par des redresseurs simple-alternances suivis par des échantillonneurs par maxima (*peak picker*). Les sorties des modèles neuronaux auxquels on réfère par motifs d'excitation auditive ont deux particularités intéressantes : ils sont positifs et épars. La partie synthèse du banc de filtres telle qu'illustrée sur la figure 3.1 doit être adaptée pour permette une synthèse parfaite. On propose dans ce chapitre une approche originale permettant de récupérer parfaitement le signal audio à partir de ses motifs d'excitation auditive. En tolérant un délai algorithmique, une simple ligne de retard peut être interposée en amont des filtres d'analyse permettant de satisfaire une qualité de synthèse parfaite validée par des métriques objectives et subjectives. Cet ensemble de lignes de retard, est largement préféré aux approches d'analyze par synthèse ou celles basées sur des processus d'approximation itératifs.

4.1 Extraction des motifs d'excitation auditives

Les modèles physiologiquement inspirés [Der *et coll.*, 2003; Erfani, 2016; Slaney, 1995] ont été utilisés auparavant pour des raisons autres que le codage de la parole et l'audio. L'objectif de ces modèles a été de comprendre la perception [Cooper, 1980; Irino et Kawahara, 1993], de tester la précision du modèle auditif [Hukin et Damper, 1989; Slaney *et coll.*, 1994] et d'améliorer les performance des algorithmes de reconnaissances de la voix. Le processus d'inversion commun de ces modèles est itératif en nature. Souvent un algorithme d'approximation itératif est utilisé pour récupérer le signal à partir des motifs d'excitation [Decorsière *et coll.*, 2015; Pichevar *et coll.*, 2010; Slaney *et coll.*, 1994].

Certaines de ces procédures d'inversion sont basées sur la théorie de projection dans des espaces convexes, où les contraintes de reconstruction sont spécifiées sous forme de minimisation convexe [Combettes, 1993]. Les projections dans ces espaces convexes pour la discipline de codage audio souffrent de plusieurs désavantages :

- La complexité élevée de cette approche due à sa nature itérative.
- Certaines contraintes ne peuvent être formulées sous forme d'une minimisation convexe.

Il est à noter que dans [Griffin et Lim, 1984], les auteurs proposent une approche itérative pour la synthèse des représentations cochléaires. Cette approche garantit une convergence optimale localement, mais encore une fois, elle requière un effort computationnel élevé. Pour ces raisons, on a choisi d'adopter une approche basée sur l'inversion des motifs d'excitation de la MB. Pour y arriver on propose d'interposer entre le bancs de filtres d'analyse et celui de synthèse des processus physiologiquement inspirés dans le but de mimer le fonctionnement du système auditif humain. Quant à la partie de synthèse on propose d'utiliser des simples gains/retards pour égaliser les sorties des filtres de synthèses. Cette approche est plus simple et ne requière nullement des processus itératifs de reconstruction.

4.1.1 Modèle des cellules ciliées internes

Les cellules ciliées internes jouent le rôle d'un transducteur qui permet de traduire les déplacements de la MB en différences de potentiel. Les mesures des réponses électriques ont montré une sensibilité directionnelle de ces capteurs : alors que le déplacement de la MB dans une direction est excitateur, le déplacement dans la direction opposée ne génère pas de décharges au niveau du NA [Dallos, 1996]. Et donc, ces cellules réagissent aux déplacements positifs de la MB et par conséquent il semble être justifié de modéliser ce comportement par un redresseur simple-alternance. Ce modèle simple a souvent été adopté pour modéliser le fonctionnement des cils ciliés internes [Baumgarte, 2001; Dau *et coll.*, 1996; Lee *et coll.*, 2015; Lyon, 1983; Seneff, 1990].

Le banc de filtres cBIT_2^* est donc suivi par un redresseur simple-alternance dont l'expression est donnée par :

$$y(n) = \max(x(n), 0)^c \tag{4.1}$$

Où $c = 0.4$ est un paramètre implémentant une compression logarithmique [Kubin et Kleijn, 1999b].

4.1.2 Modèle neuronal simple

Contrairement aux autres modèles (par example dans [Baumgarte, 2001; Dau *et coll.*, 1996; Lyon, 1983; Seneff, 1990; Thiemann, 2011; Thiemann et Kabal, 2007]), on préserve la fine

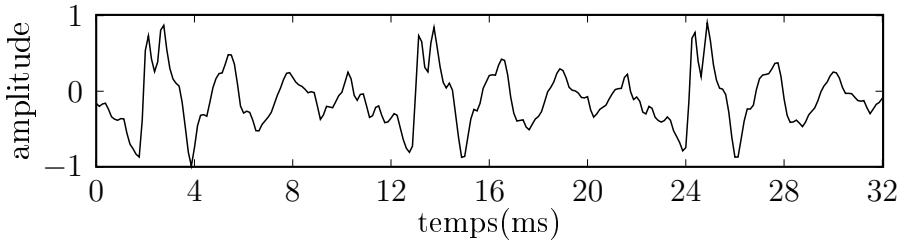
structure temporelle du signal, c'est à dire, qu'on n'applique pas de filtrage passe-bas dans le but d'extraire des enveloppes car cela conduit à une mauvaise qualité de reconstruction. Cette argumentation est basée sur les travaux de [Heinz et Swaminathan, 2009; Shamma et Lorenzi, 2013] où une étude détaillée montre l'importance de la fine structure temporelle pour l'intelligibilité des signaux de parole quand le rapport signal sur bruit est mauvais. Dans le modèle proposé, on suit le redresseur simple alternance cité plus haut par un mécanisme de sous-échantillonnage adaptatif. Ce modèle cherche les maxima locaux et met les autres échantillons à zero. Dénotant l'entrée et la sortie de ce modèle neuronal $y(n)$ et $\hat{y}(n)$ respectivement on peut alors écrire :

$$\hat{y}(n) = \begin{cases} y(n), & \text{si } y(n-1) < y(n) \text{ et } y(n+1) \leq y(n) \\ 0, & \text{sinon} \end{cases} \tag{4.2}$$

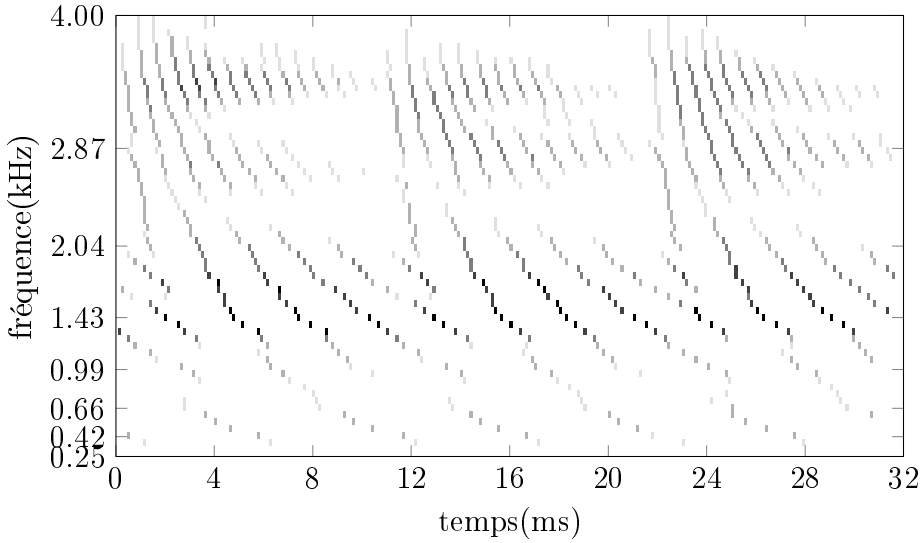
Ce modèle simule le motif de décharge d'un ensemble de neurones auditifs. Les réponses sont donc des groupes de grande activité neuronale qui sont synchronisés avec l'enveloppe du signal à l'entrée (asservies en phase). Il est connu qu'un seul neurone ne peut générer une réponse dont la fréquence dépasse les 250Hz [Greenberg, 1988; Ruggero, 1992], et donc ne peut à lui tout seul préserver la fine structure du signal à son entrée. Cependant, le system auditif humain contient bien plus de fibres nerveuses que des cils ciliés internes [Ruggero, 1992], on peut alors associer plusieurs fibres nerveuse à un seul cil cilié.

Le phénomène d'asservissement de phase ne se produisant que pour des excitations dont la fréquence est inférieure à 4 kHz [Greenberg, 1988; Li *et coll.*, 2014; Millman *et coll.*, 2015; Ruggero, 1992], l'utilisation du modèle neuronal proposé semble être justifiée pour encoder des signaux à bande étroite. Pour des raisons de simplicité, on utilisera le même modèle même pour les signaux à large bande même si cela n'est pas nécessaire. En effet, une représentation plus approximative vers les hautes fréquences est souvent suffisante (par example le cas des codeurs avec extension de bande avec et sans information additionnelle [Ekstrand, 2002; Jax et Vary, 2004; Miao *et coll.*, 2011; Valin et Lefebvre, 2000]).

La prise en compte des modèles neuronaux pulsés où l'information est véhiculée dans le timing des impulsions est clairement motivée par les observations des réseaux de neurones biologiques. Dans [Maass et Bishop, 2001; Shamma et Lorenzi, 2013], il a bien été démontré que ces modèles devraient être préférés aux modèles neuronaux classiques tels que les modèles qui opèrent en moyennant les taux de décharge au cours du temps par example. Sur la figure 4.1, le motif d'excitation généré par un segment de signal de parole voisé (phonème [ʊ] extrait du mot hib**ou**) et dont la durée est de 32 ms est donné. Pour cet



(a) Signal audio voisé.



(b) Représentation auditive d'un segment audio voisé.

Figure 4.1 Représentation auditive d'un segment audio voisé (phonème [ʊ] extrait du mot hib**ou**) d'un signal de parole. Les niveaux du gris représentent les amplitudes des sorties des modèles neuronaux.

example, 64 filtres auditifs ont été utilisés. Les sorties neuronales de ces filtres ne sont pas alignés parce que chaque filtre a un délai de groupe différent (voir la section 4.3 pour plus de détails). Cependant, le verrouillage de phase peut être clairement observé ainsi que la structure des formants autour des fréquences 1.5 kHz, 2.6 kHz et 3.4 kHz.

4.2 Synthèse par inversion des motifs d'excitation auditives

Les auteurs dans [Feldbauer, 2005] donnent une description détaillée des corrections à entreprendre avant de pouvoir synthétiser les motifs d'excitation auditives. Dans cette section, on donne un résumé de ces étapes. On couvre les deux principales causes d'une mauvaise qualité de synthèse si l'inversion des motifs d'excitation auditive se fait de façon naïve. Plus de détails peuvent être trouvés dans le chapitre 2 de [Feldbauer, 2005].

4.2.1 Inversion des modèles neuronaux

L'étape d'inversion des motifs d'excitation auditives consiste premièrement à inverser la loi de compression logarithmique implémentée suivant (4.1) :

$$z(n) = \hat{y}(n)^{1/c} \tag{4.3}$$

Où le signal $z(n)$ ressemble à un signal qui a été passé à travers un sous-échantillonneur [1] suivi d'un sur-échantillonneur opérant par insertion de zéros. Cette insertion de zéros engendre un repliement spectral (« aliasing ») auquel il faut remédier à l'aide des filtres passe-bandes localisés sur le banc de filtres de synthèse. Avant l'application de ces filtres, les amplitudes des pics doivent être corrigées pour prendre en compte la perte d'énergie. Cette perte d'énergie est engendrée par :

- le filtrage adaptatif introduit par les modèles neuronaux.
- les erreurs d'estimation de la valeur des pics en hautes fréquences dues à la fréquence d'échantillonnage finie.

Sous-échantillonnage adaptatif

À la sortie de ce modèle d'inversion, le signal ressemble donc à une sinusoïde de période $\mathbf{p}$ dont la valeur est proche de celle du filtre auditif [2]. Dans ce cas, la transformée de Fourier

1. Le terme sous-échantillonneur réfère au système décrit par l'équation (4.2)
2. Cette hypothèse n'est valide que si le signal d'entrée est riche en fréquences proches de la fréquence centrale du filtre auditif.

de ce signal est donnée par :

$$Z(e^{j\theta}) = \frac{1}{\mathbf{p}} \sum_{k=0}^{\mathbf{p}-1} z(e^{j(\theta - 2k\pi/\mathbf{p})}) \tag{4.4}$$

Applicant la transformée de Fourier à un cosinus d'amplitude unitaire et de fréquence $2\pi/\mathbf{p}$ on trouve :

$$z(e^{j\theta}) = \pi(\delta_{2\pi}(\theta - 2\pi/\mathbf{p})) + \pi(\delta_{2\pi}(\theta + 2\pi/\mathbf{p})) \tag{4.5}$$

$$Z(e^{j\theta}) = \frac{2\pi}{\mathbf{p}} \sum_{k=0}^{\mathbf{p}-1} \delta_{2\pi}(\theta - 2k\pi/\mathbf{p}) \tag{4.6}$$

Toutes les autres fréquences fantômes doivent être donc atténuées et celle du filtre auditif concerné doit être amplifiée par le facteur k_k donné par :

$$k_k = \frac{f_s}{2f_{c,k}} \tag{4.7}$$

Estimation des valeurs des pics

Les valeurs estimées de la valeur et la position du pic à la sortie du modèle neuronal ne sont pas précises. En effet, dû à une fréquence d'échantillonnage à précision finie ces erreurs sont inevitables. Pour remédier à ce problème une méthode basée sur la minimisation d'erreur quadratique moyenne a été proposée par [Kubin et Kleijn, 1999a]. Pour un sinus d'amplitude $\mathbf{A}$ et une période $\mathbf{p}$, l'amplitude maximale est observée autour de $\omega_{max} = \mathbf{A}\cos(\frac{2\pi}{\mathbf{p}})$ pour un t uniformément distribué entre $[-1/2, 1/2]$. On montre dans [Kubin et Kleijn, 1999a] qu'un estimé de la valeur de $\mathbf{A}$ est donnée par :

$$\hat{\mathbf{A}} = \beta\omega_{max} = \frac{\mathbf{p}}{\pi} \ln\left[\tan(\frac{\pi}{4 + 2\mathbf{p}})\right] \omega_{max} \tag{4.8}$$

Ce facteur de correction permet de garantir une erreur inférieure à 1dB sur toute la bande de fréquence concernée par le banc de filtres. La figure 4.2, présente la valeur du facteur de correction pour différentes valeurs de f_c/f_s.

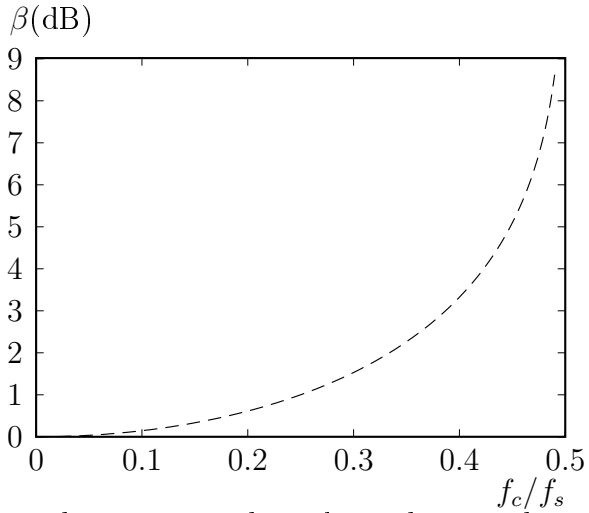


Figure 4.2 Facteur de correction des valeurs des pics du modèle neuronal.

4.3 Banc de filtres de synthèse

La dernière étape en vue d'une synthèse par inversion des motifs d'excitation auditive consiste à passer ces impulsions à travers un banc de filtres de synthèse. Pour une reconstruction « parfaite », ces filtres doivent satisfaire certaines contraintes :

- **c_1** : Posséder des caractéristiques fréquentielles passe-bande pour éliminer le phénomène de repliement spectral et limiter l'étalement spectral du bruit de quantification.
- **c_2** : Posséder un délai minimal pour que la qualité de reconstruction soit la plus proche de celle du signal d'origine.

Généralement, le banc de filtre de synthèse n'est pas unique [Kubin et Kleijn, 1999b]. Pour un banc de filtres non-décimés, la condition de reconstruction parfaite, en se permettant un délai d s'écrit :

$$G_i(z) = z^{-d} \frac{H_i(z^{-1})}{\sum_{k=1}^{n} H_i(z) H_k(z^{-1})} \tag{4.9}$$

Dans le cas où le dénominateur dans l'équation (4.9) ne vaut pas un, le banc de filtres de synthèse est le même que celui du banc d'analyze mais avec ses RIs inversées dans le temps. Pour maintenir la causalité du système d'analyse par synthèse, cette condition se traduit par un ajout de délai égal à la longueur du filtre de synthèse (Dans le cas où celui-ci est un RIF). Dans le cas contraire, par example quand le nombre des filtres d'analyse est peu élevé ou quand les filtres sont des filtres à réponse impulsionnelle infinie (RII), un filtre d'égalisation peut être introduit pour réduire les ondulations quand la synthèse se fait par simple sommation des sorties des filtres d'analyse [Foster et Herley, 1995]. Cette approche est utilisée dans les systèmes proposés par [Irino et Unoki, 1998; Lin *et coll.*, 2001; Pichevar *et coll.*, 2004; Thiemann, 2011]. Un inconvenient majeur de l'approche basée sur l'équation (4.9) est qu'il faut un délai additionnel pour que la réponse du système soit causale. Ceci

a pour effet de limiter l'utilisation de ce banc de filtres dans le cadre du codage audio temps-réel. On propose dans les sections suivantes une approche originale pour concevoir le banc de synthèse tout en maintenant un délai minimal du système. On utilise les mêmes filtres d'analyse aussi pour faire la synthèse sans inverser leurs RIs. Pour égaliser la réponse du système on introduit une ligne de retard avec des gains de compensation.

4.3.1 Synthèse sans modèles neuronaux

Commençons par considérer le cas le plus simple à savoir qu'il n'y a pas de décimation à la sortie du banc de filtres d'analyze (c'est à dire que $\hat{y}_i(n) = y_i(n)$ sur la figure 4.3). Simplifions encore le problème et imposons aux filtres de synthèse une structure simple, c'est à dire que :

$$G_i = g_i z^{-\delta_i} \quad \forall i \in [1,\ N] \tag{4.10}$$

En supposant une impulsion de Kronecker [3] à l'entrée du système, une condition nécessaire et suffisante pour garantir (4.9) est donnée par :

$$\begin{aligned} \hat{s}(k) &\approx s(k-d) \\ \hat{s}(k) &\approx \delta(k-d) \\ \sum_{i=0}^{n} g_i H_i(k-\delta_i) &\approx \delta(k-d) \end{aligned} \tag{4.11}$$

On propose dans ce qui suit de trouver les valeurs de (g_i, δ_i) qui permettent de satisfaire l'équation (4.11). C'est un problème de minimisation quadratique dont la solution peut être approchée en deux étapes.

Détermination des valeurs des délais δ_i :

Pour un délai d donné et des gains $g_i = 1$ donnés, pour que la réponse du système ressemble le plus possible à une impulsion de Kronecker, les RIs des filtres de synthèse doivent atteindre leurs valeurs maximales à l'instant d. Si cette dernière condition est satisfaite, et étant donné que ces filtres ont des fréquences centrales différentes, les RIs vont s'additionner positivement à l'instant d alors qu'elles vont partiellement s'annuler ailleurs. En effet si on définit $\delta_k^{\max}$ l'instant auquel la RI du filtre k atteint sa valeur maximale entre [0, d], on peut écrire :

$$\delta_k = d - \delta_k^{\max} \tag{4.12}$$

3. Cette impulsion vaut 1 à l'origine et 0 ailleurs.

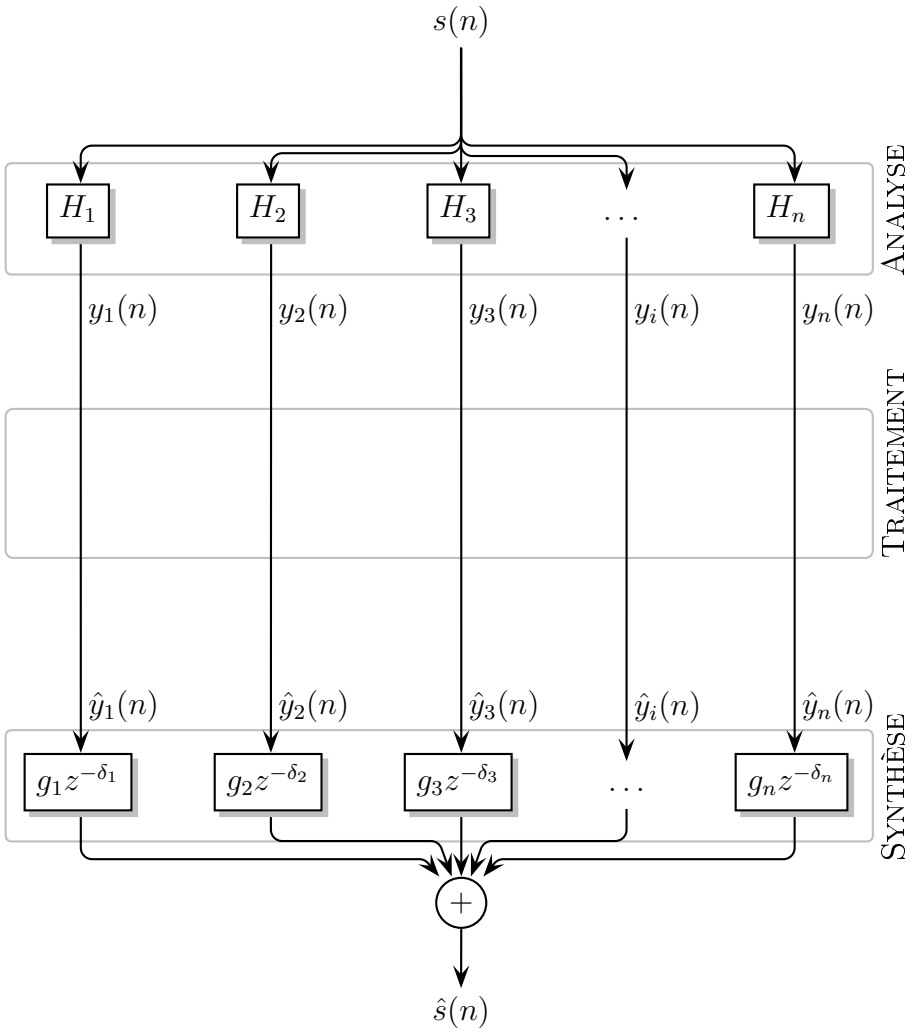


Figure 4.3 Structure en parallèle du filtre auditif proposé.

La figure 4.4 donne un example de l'application du délai déterminé par l'équation (4.12). Si on considère un banc de filtres à n canaux et si on définit les filtres de synthèse comme étant $G_i = z^{-\delta_i} \forall i \in [1,\, n]$ où les δ_i sont déterminés comme décrit précédemment, la RI $s(n)$ du système est donnée par :

$$s(n) = \sum_{k=1}^{N} h_k(n - \delta_k) \tag{4.13}$$

La figure 4.5 présente la réponse d'un tel système d'analyse-synthèse. La figure 4.5(a) présente la réponse impulsionnelle du système alors que la figure 4.5(b) présente sa réponse fréquentielle pour un délai d=4ms et des gains unitaires. La RI du système atteint son maximum à l'instant d mais des maximaux locaux sont visibles sur la RI de la figure 4.5(a). Ceci se traduit par une réponse fréquentielle partiellement plate (idéalement elle

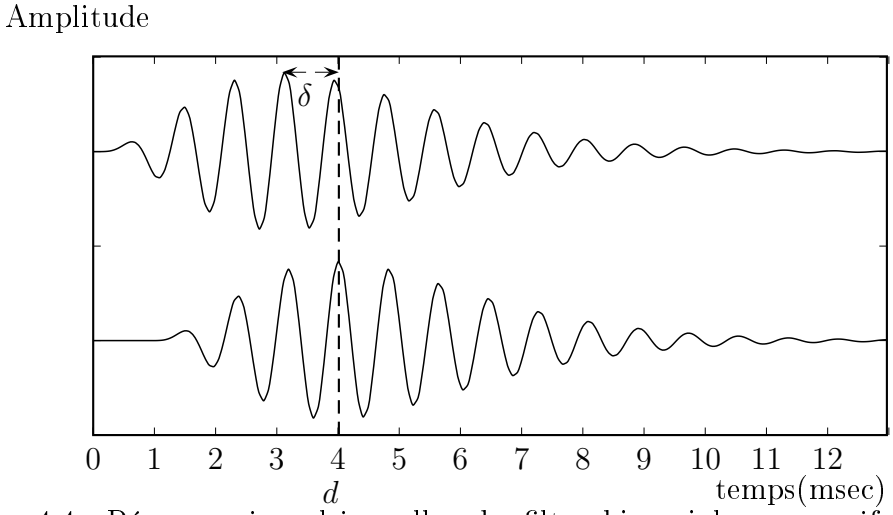
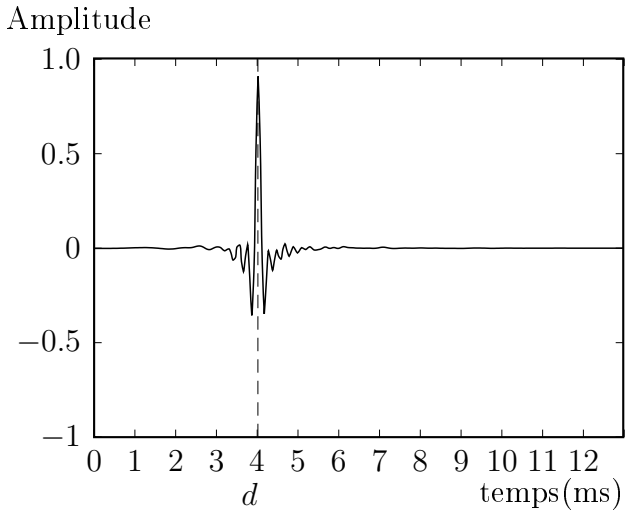
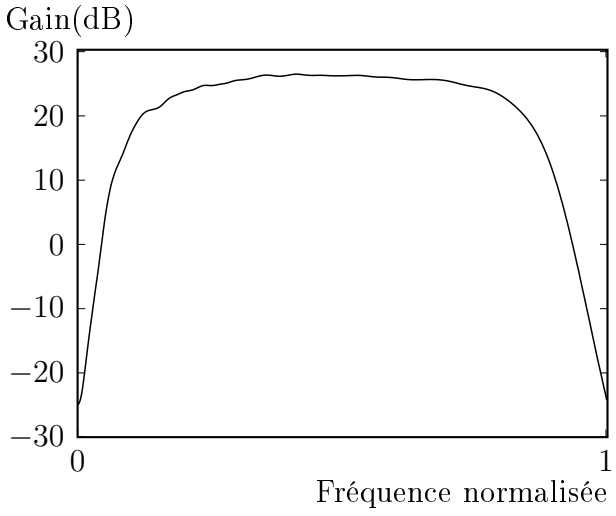


Figure 4.4 Réponses impulsionnelles du filtre binomial compressif avant et après l'insertion du délai δ pour que celle-ci atteigne sa valeur maximale exactement à l'instant $d = 4$ms.



(a) Réponse impulsionnelle du banc de filtres d'analyse-synthèse.



(b) Réponse fréquentielle du banc de filtres d'analyse-synthèse.

Figure 4.5 Réponse impulsionnelle et fréquentielle du banc d'analyse-synthèse dont le banc de synthèse est décrit par l'équation (4.10) avec des gains g_i unitaires.

devrait être complètement unitaire pour toutes les fréquences). La sous-section suivante a pour but de donner une méthode pour trouver les valeurs des gains g_i qui permettent de vérifier partiellement la condition de reconstruction parfaite.

Détermination des valeurs des gains g_i :

On rappelle l'expression de la RI du système d'analyse-synthèse :

$$\delta_{n-d} \approx \sum_{k=1}^{N} g_k h_k(n - \delta_k) \tag{4.14}$$

avec les délais δ_k déterminés comme décrit précédemment. L'équation (4.14) peut être traduite en une approximation linéaire écrite sous forme matricielle :

$$\begin{pmatrix} 0 & \dots & 0 \\ \vdots & h_k(n) & 0 \\ h_1(n) & h_k(n-1) & h_N(n) \\ \vdots & \vdots & \vdots \\ h_1(n-M-\delta_1) & h_k(n-M-\delta_k) & h_N(n-M-\delta_N) \end{pmatrix} \times \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{N-1} \\ g_N \end{pmatrix} \approx \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \tag{4.15}$$

$$\longleftrightarrow$$

$$\mathbf{H} \times \mathbf{g} \approx \delta_{n-d} \tag{4.16}$$

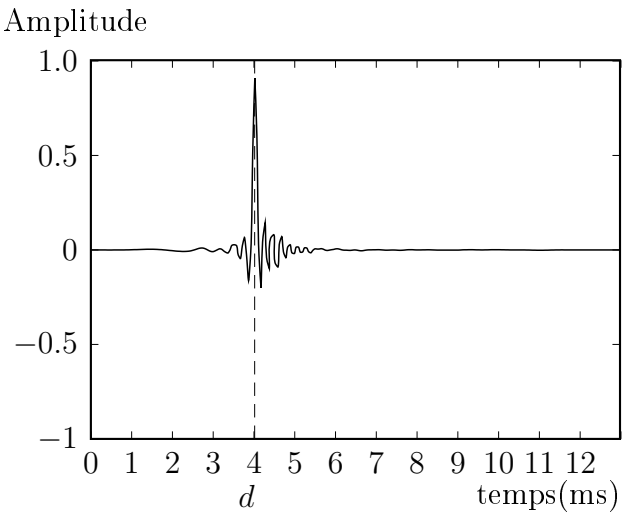
Où $\mathbf{H} \in \mathbb{R}_{M,N}$, $\mathbf{g} \in \mathbb{R}_{N,1}$ et :

$$\mathbf{H}_{i,j} = \begin{cases} h_i(n + \delta_i - j) \text{ si } j \geq \delta_i \\ 0 \text{ sinon} \end{cases} \tag{4.17}$$

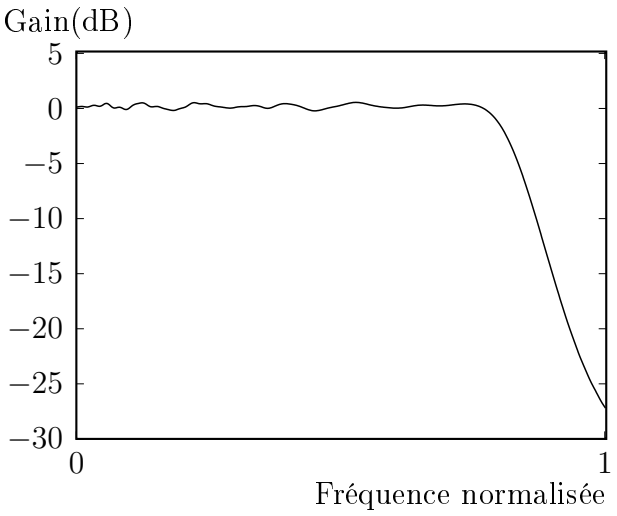
La solution de l'équation sous-déterminée (4.16) peut être trouvée par pseudo-inversion :

$$\mathbf{g} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \times \delta_{n-d} \tag{4.18}$$

Où $(.)^T$ est l'opération de transposition matricielle. La figure 4.6 donne le résultat de l'application des gains déterminés par l'équation (4.18). Dans cet example, 64 filtres ont été utilisés et l'équation sous-déterminée (4.18) est résolue numériquement utilisant des RIs dont les longueurs sont de 100ms. Ces longueurs permettent de tenir compte de l'étalement de la RI du filtre auditif dont la fréquence centrale est la plus petite (50Hz). La RI ressemble plus à une impulsion de Kronecker ce qui se traduit par une réponse fréquentielle plate sur la figure 4.6(b). Dans l'example présenté sur la figure citée ci-haut, la variation de la réponse fréquentielle est inférieure à 0.5dB sur toute la bande fréquentielle concernée. Il est évident que les qualités objectives et subjectives de synthèse dépendent principalement de deux facteurs principaux à savoir N le nombre de filtres ainsi que le délai d.



(a) Réponse impulsionnelle du banc de filtres d'analyse-synthèse.



(b) Réponse fréquentielle du banc de filtres d'analyse-synthèse.

Figure 4.6 Réponse impulsionnelle et fréquentielle du banc d'analyse-synthèse dont le banc de synthèse est décrit par l'équation (4.10) avec des gains et délais déterminés par les équations (4.18) et (4.12).

Avant de passer à la determination des paramètres optimaux du banc de filtres de synthèse, il est à noter que le système donné par l'équation (4.14) peut être écrit dans le domaine de la transformée de Fourier. Il suffit d'appliquer la transformée de Fourier à l'équation (4.14) ce qui donne :

$$1 \approx \sum_{k=1}^{N} g_k H_k(f) \ \forall f \in [0,\ f_s/2] \tag{4.19}$$

Cette dernière équation discrètisée peut s'écrire sous forme matricielle :

$$\begin{pmatrix} H_1(0) & H_2(0) & H_N(0) \\ H_1(f_1) & \vdots & H_N(f_1) \\ \vdots & H_2(f_{n-1}) & \vdots \\ H_1(f_n) & H_2(f_n) & H_N(f_n) \end{pmatrix} \times \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{N-1} \\ g_N \end{pmatrix} \approx \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{4.20}$$

$$\longleftrightarrow$$

$$\mathbf{H_f} \times \mathbf{g} \approx I_{N,1} \tag{4.21}$$

Où $H_i(f_k)$ est la réponse fréquentielle du filtre h_k estimée à la fréquence f_k et $I_{N,1}$ est le vecteur valant 1 partout. Pareillement à l'équation (4.18), la valeur de $\mathbf{g}$ peut être trouvée

par pseudo-inversion :

$$\mathbf{g} = (\mathbf{H_f}^T \mathbf{H_f})^{-1} \mathbf{H_f}^T \times I_{N,1} \tag{4.22}$$

Évidement les valeurs données par l'équation (4.18) différent de celles données par l'équation (4.22) et cela conduit nécessairement à deux systèmes dont les performances diffèrent.

Résultats de simulation :

On présente dans cette section le résultat de passage d'une trame d'un signal de parole échantillonné à 16 kHz à travers le bancs de filtres analyse-synthèse. On a utilisé un banc de filtres cBIT_2^* contenant 64 filtres placés de façon uniforme sur l'échelle ERB et couvrant la bande de fréquence [50Hz, 7.5Khz]. Pour cette expérience le délai d a été fixé à 25ms. La figure présente les résultats obtenus. Les gains trouvés par la méthode d'égalisation

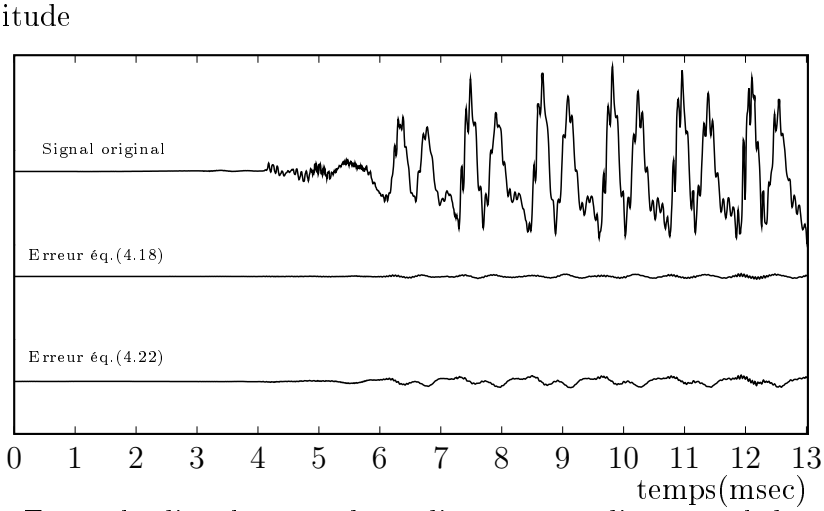


Figure 4.7 Example d'analyse-synthèse d'une trame d'un signal de parole et les erreurs de synthèse encourues pour différentes méthodes d'optimisation de gains.

dans le domaine fréquentiel (g_i) donnent des moins bons résultats que ceux trouvés par la méthode d'égalisation dans le domaine temporel. En effet le RSB du système avec des gains déterminés selon l'équation (4.18) vaut 32 dB contre seulement 22 dB pour des gains trouvés selon l'équation (4.22). Ceci s'explique par le fait que l'équation (4.18) prend en compte l'imperfection de la RI du système (figure 4.6(a)) puisqu'elle agit directement sur celle-ci alors que (4.22) suppose une RI parfaite et opère sur une version discrètisée de celle-ci. Les résultats présentés dans cette sous-section montrent que cette approche d'égalisation est suffisante dans le cas où le système d'analyse-synthèse opère en sommant des versions décalées des sorties du banc de filtres de synthèse. Dans la sous-section suivante, on aborde le cas où les sorties de chaque canal est passée à travers un modèle neuronal décrit par l'équation (4.2).

4.3.2 Synthèse avec intégration du modèle neuronal

Comme expliqué dans la section 4.2.1, à la sortie du modèle neuronal opérant sur le filtre d'analyse de fréquence centrale f_k, le signal ressemble à un signal sous-échantillonné ce qui crée forcément des composantes spectrales indésirables. Dans [Feldbauer, 2005; Feldbauer et Kubin, 2004], les auteurs proposent des filtres de synthèse dont les RIs sont celles des filtres d'analyse (GT) mais inversées dans le temps (voir équation (4.9)). Utilisant des filtres GTs implémentés sous forme de filtres RIFs d'ordre 666, pour une fréquence d'échantillonnage de 8Khz le délai de leur système de 20 canaux est de 83.25 ms (section 2.3.3 de [Feldbauer, 2005]). Les auteurs justifient leurs approche par la théorie des trames expliquée dans la section 2.3.4 de [Feldbauer, 2005]. Le même modèle est aussi utilisé dans [Thiemann, 2011]. Pour la même configuration, l'utilisation du banc de filtres proposés dans cette thèse réduirait la complexité d'implémentation d'extraction et inversion des motifs auditif par un ordre de 1/40.

On propose une approche différente que celle proposée dans [Feldbauer, 2005] où celle proposée dans [Thiemann, 2011]. On rappelle qu'une synthèse parfaite dans le cas d'un système décimé contraint les filtres de synthèse à posséder une caractéristique passe-bande. On a choisit d'utiliser les mêmes filtres d'analyse pour l'étape de synthèse puisque ceux-ci ont des caractéristiques passe-bande ce qui permet d'éliminer le repliement spectral dû au sous-échantillonnage adaptatif introduit par les modèles neuronaux (voir section 4.2.1). On utilise la même méthode décrite par l'équation (4.16) pour égaliser la RI du système d'analyse-synthèse. On note que dans cette approche il n'est plus nécessaire d'inclure les étapes de correction détaillées dans 4.2.1. En effet, les gains g_i trouvés avec l'équation (4.18) englobent les facteurs de correction donnés dans les équations (4.7) et (4.8).

L'approche proposée est différente de celle utilisée par [Feldbauer, 2005; Irino et Unoki, 1998; Pichevar *et coll.*, 2004; Thiemann, 2011] :

- Les filtres utilisés sont des BITs et non des filtres GTs.
- Les RIs des filtres de synthèse sont les mêmes que celles des filtres d'analyse et non inversées dans le temps. Ceci a pour effet d'aboutir à un système d'analyse-synthèse à moindre délai algorithmique.
- Pour compenser la réponse du système, on utilise une ligne de compensation de gains et une simple ligne de retard au lieu d'utiliser un égaliseur comme utilisé dans [Feldbauer, 2005].

La figure 4.8 présente la nouvelle architecture du système proposée où les modèles neuronaux ont été représentés par des sous-échantillonneurs adaptatifs. Pour une valeur de délai d donné et un nombre de filtres N, la determination des valeurs des gains g_i et délais δ_i se

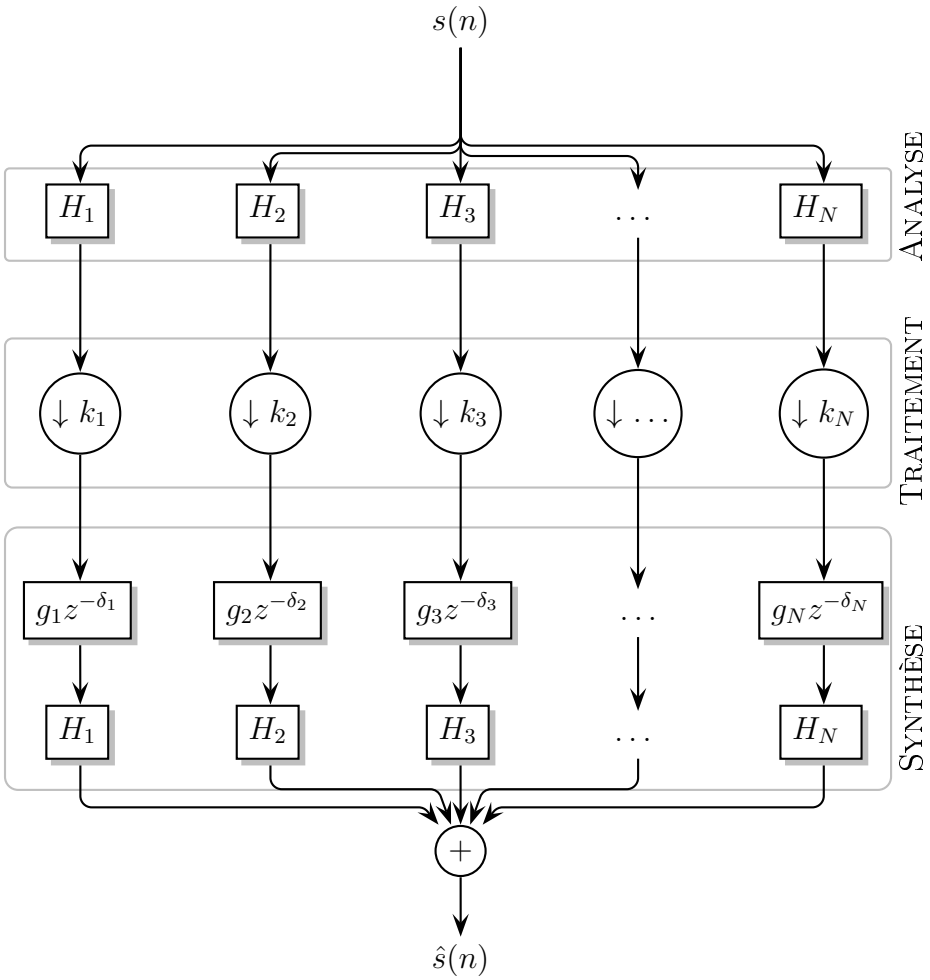


Figure 4.8 Structure en parallèle du filtre auditif proposé incluant les modèles neuronaux.

fait par la résolution de l'équation (4.18) et l'équation (4.12). Dans la section suivante on propose d'évaluer les performances du système proposé en terme de qualité de synthèse.

4.4 Résultats expérimentaux

On propose dans cette section d'évaluer les qualités du système proposé. Pour ce faire, deux métriques sont utilisées : le RSB et la différence de qualité subjective, *subjective difference grade* (SDG). Alors que le RSB peut être calculé facilement à partir des signaux audio (référence et dégradé), la SDG nécessite la réalisation des tests d'écoute qui est une tache souvent fastidieuse, longue et coûteuse.

Plusieurs algorithmes ont été développés pour prédire l'issue d'un test d'écoute effectué selon la recommendation [ITU-BS-1116, 1997]. Durant ce test, la différence entre deux

stimuli, l'élément à juger et la référence cachée sont notés sur une échelle allant de 5 à 1 représentant respectivement une dégradation « imperceptible » à « très agaçante ». En supposant que les auditeurs attribuent une meilleure note à la référence cachée qu'à l'élément sous test, la différence entre la note de la référence et ce dernier est comprise entre 0 et -4 et appelée SDG. L'algorithme de l'évaluation perceptuelle de la qualité audio, *Perceptual evaluation of audio quality* (PEAQ) proposé par [ITU-BS-1387, 2001] calcule la différence de qualité objective, *objective difference grade* (ODG) qui est destinée à ressembler à la SDG. L'implémentation de cet algorithme étant protégée par des droits d'auteurs, des implémentations « open source » basées sur la recommendation [ITU-BS-1116, 1997] ont été développées.

Comme mentionné par [Kabal, 2002], la recommendation [ITU-BS-1387, 2001] n'est pas assez spécifique et les données ne sont pas suffisantes pour complètement caractériser l'algorithme PEAQ. Ceci explique pourquoi seule Opticom [OPTICOM, 2016] (activement impliquée dans la mise au point de cette recommendation) possède une implémentation propriétaire conforme à la recommendation [ITU-BS-1387, 2001]. Dans [Kabal, 2002], les auteurs fournissent une implémentation MatLab de cet algorithme qui n'est pas conforme à la recommendation mais qui s'en approche. Dans [Holters et Zölzer, 2015] les auteurs présentent le *GstPeaq* : une implémentation du même algorithme mais qui valide de façon plus proche la dite recommendation. La déviation entre les scores ODG fournis par *GstPeaq* et la recommendation pour les signaux de test est inférieure à 0.18. Le code source du *GstPeaq*[4] a été télécopié depuis le site[5] publié dans [Holters et Zölzer, 2015] et compilé sur une machine Linux sans aucune modification. La figure 4.9 donne la valeur ODG en fonction de la valeur de la dégradation subjective (SDG). Cette figure montre la forte corrélation entre les deux métriques. Il est donc possible d'obtenir une estimation fiable de la qualité de synthèse d'un codeur audio en utilisant l'ODG. Cette estimation est encore plus fiable pour des valeurs ODG supérieures à -1.5 [ITU-BS-1387, 2001].

Pour estimer la qualité du système d'analyse-synthèse proposé, des signaux tests sont utilisés pour calculer le RSB et le ODG pour différentes combinaisons des paramètres (N : nombre de filtres et d délai du système) du système proposé. Ces signaux échantillonnés à une fréquence de 16 kHz contiennent des signaux de parole ainsi que des signaux de musique riches en harmoniques (harpe, violon, triangle etc) et percussions (musique d'orchestre, castagnettes). Une description détaillé du contenu de ces signaux est donnée [ITU-BS-1387, 2015] section 7.3. Pour un nombre de filtres N donné, le banc de filtres d'analyse-synthèse

4. Dans la suite du document on va référencer l'implémentation *GstPeaq* comme étant le PEAQ.
5. `https://github.com/HSU-ANT/gstpeaq` consulté le 20-06-2015.

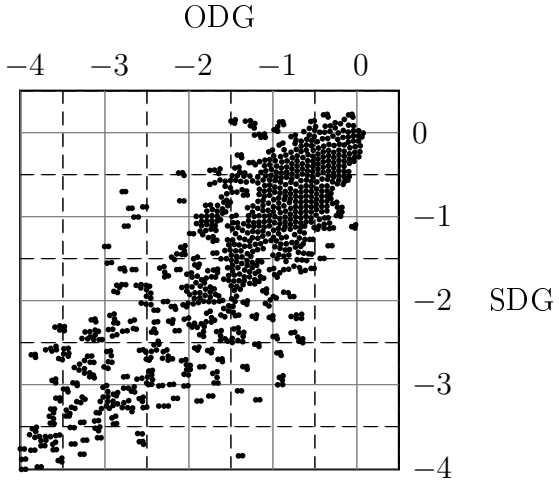


Figure 4.9 Relation entre l'ODG et le SDG.

est conçu chaque fois pour couvrir la bande [50 Hz, 7.5 Khz] de façon uniforme sur l'échelle ERB.

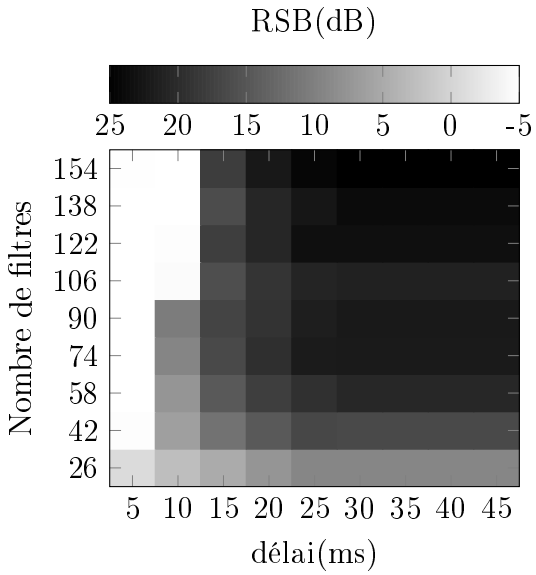
Le choix de la fréquence d'échantillonnage est dicté par la fréquence maximale utilisée pour l'ajustement du filtre cBIT_2^*. En effet, la fréquence maximale des expériences de masquage dont on dispose ne couvre que la bande [250 Hz, 6 kHz] (voir section 3.2.1).

Dans cette expérience, le nombre de filtres a été varié entre 26 et 154 (correspondant à un nombre de filtres par ERB variant entre 0.8 et 5) alors que le délai d a été varié entre 5 et 40 ms. La figure 4.10 présente les résultats de cette simulation.

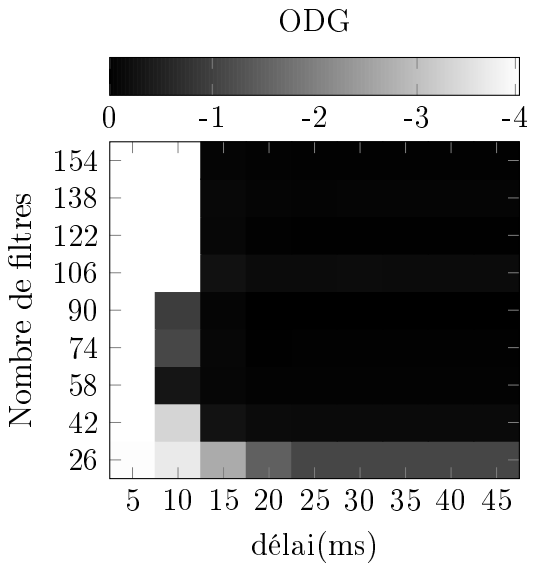
La figure 4.10(a) donne le RSB moyen entre les signaux de test et les signaux synthétisés. Pour des délais inférieurs à 10 ms, le fait d'augmenter le nombre de filtres ne permet pas d'améliorer la qualité de synthèse. En effet, pour de si petits délais, les valeurs maximales des RIs se produisent (pour les basses fréquences) après le délai d et donc une fois alignés, ces RIs ne s'annulent pas parfaitement pour ressembler à une impulsion de Kronecker.

Pour des délais supérieurs à 15 ms et pour un nombre de filtres N supérieur à 74, la qualité de synthèse est bonne. Cela se traduit par des valeurs de RSB supérieures à 10 dB et des valeurs de ODG supérieures à -1 suggérant une dégradation imperceptible ou perceptible mais non « agaçante ».

Pour un délai de 25 ms et un nombre de filtres de 74 (2.4 filtres par ERB), le RSB moyen et le ODG valent 24±2dB et -0.07±0.08 respectivement. Avec ces valeurs, la dégradation moyenne entre analyse est synthèse est imperceptible. Un délai algorithmique de 25 ms n'est pas une valeur aberrante pour un codec audio large bande opérant à une fréquence



(a) RSB moyen entre références et signaux synthétisés.



(b) ODG moyen entre références et signaux synthétisés.

Figure 4.10 RSB et ODG moyens entre références et signaux synthétisés à partir de leurs motifs d'excitation auditive pour différents paramètres du système analyse-synthèse.

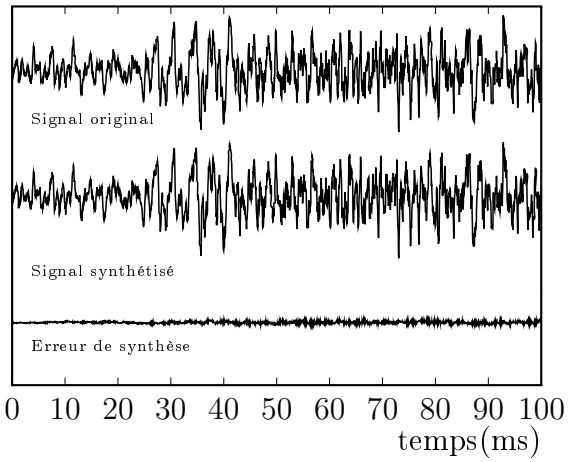
d'échantillonnage de 16 kHz. Dans [Lutzky *et coll.*, 2004], les auteurs présentent une revue des sources de délais que des codecs audio communs introduisent lors la compression de signaux audio. Ces délais algorithmiques [6] varient entre 20 ms pour MEPG-4 AAC-LowDelay [Brandenburg et Bosi, 1997] et 25 ms pour le codec AMR-WB [3GPP, 2007]. Pour d'autres codecs, par exemple le MPEG-HE AAC [Wolters *et coll.*, 2003], cela peut aller jusqu'à 129 ms.

La figure 4.11 présente deux examples de synthèse de signaux audio ainsi que l'erreur de synthèse. L'algorithme proposé permet de synthétiser « parfaitement » [7] les signaux audio à partir de leurs motifs d'excitation auditive (figure 4.11(d) et figure 4.11(c)). Le RSB vaut 27 dB et 28 dB pour ces trames de musique et de parole respectivement.

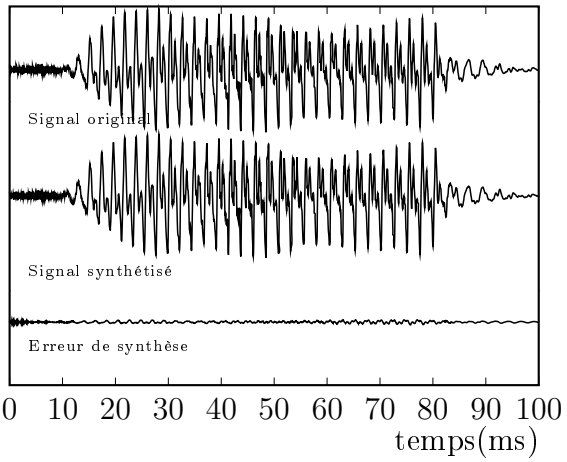
Malgré que le modèle neuronal utilisé réduit approximativement par moitié le nombre de pics présents sur le motif d'excitation le rendant ainsi épars, la transmission de telle information nécessiterait un débit très élevé. Définissant τ_s rapport de parcimonie (*sparsity*) comme étant le rapport entre le nombre d'éléments nuls sur le nombre total d'éléments

6. Ces délais excluent toute prise en compte de l'implémentation et supposent une implémentation sur un système où les ressources computationnelles sont illimitées.

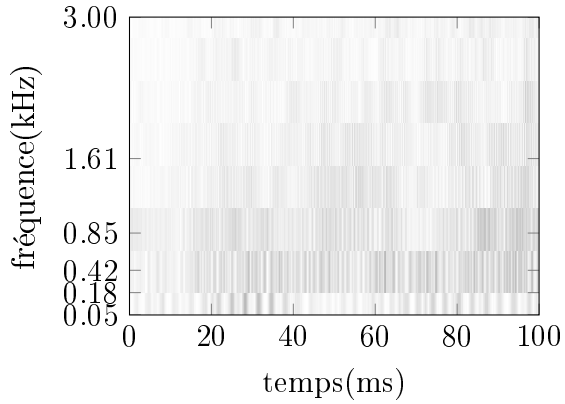
7. L'erreur de synthèse est majoritairement due à l'absence de filtres vers les très basses fréquences et les très hautes fréquences.



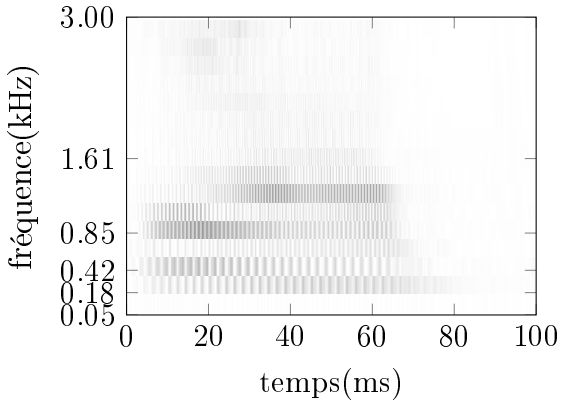
(a) Example de synthèse d'un signal de musique à partir de son motif d'excitation auditive.



(b) Example de synthèse d'un signal de parole à partir de son motif d'excitation auditive.



(c) Motif d'excitation auditive d'un signal de musique.



(d) Motif d'excitation auditive d'un signal de parole.

Figure 4.11 Example de synthèse de signaux à partir de leurs motifs d'excitation auditive. La fréquence d'échantillonnage est de 16 kHz mais pour des raisons de clarté, l'axe vertical de ces motifs a été limité à 3 kHz.

non nuls d'un signal x, on peut estimer de façon optimiste le débit moyen comme étant :

$$\text{Débit}(x) = \delta_s(x)\mathbb{H}(x \neq 0) \tag{4.23}$$

Où $\mathbb{H}$ est l'entropie donnée en bits/seconde.

Pour avoir une idée sur l'ordre de grandeur de cette quantité, on a réalisé une simulation avec plusieurs signaux audio. Utilisant l'équation (4.23), cette quantité vaut 3.6 Mbit/sec et 1.4 Mbit/sec pour des signaux de musique et de parole respectivement. Le chapitre suivant présente les algorithmes de compression avec et sans perte permettant de réduire cette quantité tout en préservant une bonne qualité à l'écoute.

4.5 Discussions

Dans [Thiemann et Kabal, 2007] et [Thiemann, 2011] le modèle proposé par [Feldbauer, 2005] a été modifié en remplaçant le modèle neuronal par des détecteurs d'enveloppe basés sur la transformée de Hilbert. Alors que dans cette thèse l'analyse se fait par inversion des motifs d'excitation auditive, dans les travaux cités plus haut, l'analyse se fait par approche itérative où le signal original est récupéré comme étant une somme pondérés des enveloppes filtrées par le banc de synthèse. Outre la complexité d'implémentation inhérente à une approche d'analyze par synthèse, la convergence vers un minimum de reconstruction global n'est pas toujours garantie. Dans le chapitre 5 du même ouvrage, Thiemann [2011] donne les résultats de simulation de l'approche proposée. Même sans inclusion de modèles de masquage, la qualité de synthèse n'est pas bonne (le signal ancre [8] score significativement mieux que le signal reconstruit à partir de sa représentation complète). L'auteur montre par la même occasion que les performances sont très dépendantes du signal analysé : pour la même configuration du système, la qualité de synthèse diffère grandement en fonction du signal analysé. Sur la figure 5.8 de [Thiemann, 2011] un example de deux signaux est donné où un des signaux a un score MUSHRA double de l'autre pour la même configuration du système. Même si dans [Decorsière *et coll.*, 2015] une approche plus sophistiquée est suivie pour la synthèse à partir des enveloppes, le problème de la complexité d'une telle approche reste encore posé. Par example il faut en moyenne 80 itérations par trame pour obtenir une bonne qualité de synthèse [Decorsière *et coll.*, 2015].

Dans [Pichevar *et coll.*, 2010] les auteurs proposent une approche pour la décomposition d'un signal audio en un ensemble d'objets sonores en utilisant une base sur-complète

8. L'ancre est le signal de référence dont la largeur de bande est limitée à 4 kHz. Dans les tests de type MUSHRA, ce signal fourni une référence qui devrait être la moins bien notée.

d'atomes à base de filtres GCs. Cette piste de recherche a été exclue dû à sa nature itérative puisqu'elle se base sur l'algorithme du *matching pursuit* (MP). Même si les auteurs proposent une approche moins complexe en terme d'implémentation [Pichevar *et coll.*, 2011], on pense que cette approche itérative reste complexe en terme d'implémentation ce qui la rend moins appropriée pour le codage temps-réel. Il faut en moyenne 2000 itérations pour que ces algorithmes convergent. Il nous a aussi semblé que les approches par projection sur des bases sur-complètes ne trouvent pas facilement une justification intuitive de point de vue fonctionnement biologique du système auditif.

4.6 Conclusion

Dans ce chapitre, on a présenté les algorithmes qui permettent de transformer un signal audio en motifs d'excitation auditive. Chaque filtre cBIT_2^* suivi d'un modèle neuronal simple permet de produire un train d'impulsions mimant la réponse du système auditif humain pour une position donnée sur la cochlée. Pour synthétiser le signal audio à partir de ces trains d'impulsions, le banc de filtres de synthèse opère en sommant des versions décalées et amplifiées de ses dernières passées à travers le même banc de filtres d'analyse. On a montré que cette approche d'analyze-synthèse permet de recouvrir sans pertes audibles le signal original tout en introduisant un délai minimal. Cette approche est beaucoup moins complexe en implémentation que celles proposées dans la littérature. Par example on réduit par un facteur de 1/40 la complexité d'implémentation de l'extraction et l'inversion des motifs de l'excitation auditive comparativement à [Thiemann, 2011]. Cependant, la transmission de ces derniers tels qu'extraits et décrits dans la section 4.3 requerrait un débit très élevé. Le chapitre suivant présente des algorithmes de masquage opérant dans le domaine perceptuel permettant de réduire le nombre d'impulsions tout en maintenant une bonne qualité de synthèse.

CHAPITRE 5

Masquage dans le domaine perceptuel

Dans le chapitre précédant, on a décrit les étapes entreprises pour extraire les motifs d'excitation auditive d'un signal audio. On a aussi montré qu'avec le bon choix des paramètres de ce modèle, on est capable de resynthétiser fidèlement le signal d'origine par inversion de ses motifs d'excitation auditive. Les blocs utilisés étant biologiquement inspirés, cette approche peut être considérée comme étant une base de projection d'un signal audio dans le domaine perceptuel. Ceci dit, cette projection est très redondante. En effet, le nombre d'impulsions générées par cette projection est linéairement proportionnel à la longueur du signal audio multipliée par le nombre de filtres auditifs. Ce chapitre décrit les approches possibles pour réduire le nombre d'impulsions produits en exploitant les limitations du système auditif humain. On introduit dans ce chapitre un nouvel algorithme de masquage opérant dans le domaine des motifs d'excitation auditive permettant de réduire le nombre d'impulsions tout en maintenant une bonne qualité de synthèse. L'estimation de la qualité de ce nouvel algorithme est conduite utilisant une variété de signaux de parole : Pour un nombre d'impulsions par échantillon aussi petit que 0.76 la qualité de synthèse est bonne et est confirmée par une valeur d'ODG moyen valant -1.25.

5.1 Masquage et parcimonie

Les motifs auditifs extraits par le modèle décrit dans le chapitre précédant sont épars contenant un nombre limité d'éléments non nuls. Cependant, ces motifs contiennent plus d'éléments non nuls que le signal original. Plusieurs algorithmes ont été développés pour réduire le nombre d'impulsions. Ces algorithmes peuvent être classés en masquage *simultané* et masquage *temporel*. Dans [Allen, 2008; Brandenburg, 1999; Kubin et Kleijn, 1999b] il a été montré que le masquage temporel réduit bien plus le nombre d'impulsions que le masquage simultané.

Dans [Kubin et Kleijn, 1999b], un modèle simple de masquage post-stimuli « forward masking » a été utilisé. Si on suppose que le train d'impulsions pour un canal donné est

noté $x(n)$ et que le seuil du masquage post-stimuli est noté $T(n)$, alors :

$$T(n) = \begin{cases} x(n) \text{ si } x(n) \geq T(n-1)e^{-1/\tau} \\ T(n-1)e^{-1/\tau} \text{ sinon} \end{cases} \tag{5.1}$$

Où τ est une constante de temps dépendante de la fréquence centrale du filtre considéré et est déterminée empiriquement [Kubin et Kleijn, 1999b]. Une fois le seuil $T(n)$ calculé, le signal à la sortie de l'étape du masquage, $y(n)$ est donné par :

$$y(n) = \begin{cases} x(n) \text{ si } x(n) \geq T(n-1)e^{-1/\tau} \\ 0 \text{ sinon} \end{cases} \tag{5.2}$$

Les auteurs prétendent qu'ils sont parvenus à réduire de moitié le nombre d'impulsions pour un signal de parole échantillonné à 16kHz sans une dégradation notable de la qualité à l'écoute.

Dans [Feldbauer, 2005] des motifs précalculés d'excitation unitaire sont utilisés pour déterminer les seuils de masquage. Chaque canal de leur banc de filtres GTs est excité par une impulsion de Dirac ensuite l'enveloppe de Hilbert est calculée. Cela permet d'avoir ce que les auteurs appellent un motif d'excitation d'impulsion isolée (*Isolated-pulse BM excitation pattern*). Plus précisément, chaque RI du filtre d'analyse est convoluée avec les RIs de tous les filtres de synthèse, ce qui permet d'avoir une représentation en deux dimensions (temps, fréquence) de l'effet d'une impulsion à l'entrée de leur banc de filtres. L'enveloppe de cette représentation est ensuite extraite et est notée comme étant motif d'excitation d'impulsion isolée. Si on note l'indice du filtre d'intérêt ch alors ce motif d'excitation isolé à l'instant n et à la fréquence indexée k, est donné par :

$$E_{ch}(n,k) = \mathcal{E}(g_{ch}(n) * h_k(n)) \tag{5.3}$$

Où $\mathcal{E}$ représente l'opération d'extraction d'enveloppe utilisant la transformée de Hilbert $\mathcal{H}$ donnée par :

$$\mathcal{E}(f) = |f + j\mathcal{H}(f)| \tag{5.4}$$

Le critère de masquage proposé par [Feldbauer, 2005] considère que les impulsions ayant la plus grande amplitude ont un pouvoir masquant plus important (véhicule une grande partie d'information contenue dans le motif d'excitation auditive) et donc cette approche est une approche itérative. Pour chaque trame, les impulsions sont triées selon leurs amplitudes. Chaque impulsion est alors considérée comme une impulsion masquante, les autres

impulsions étant des candidates pour être masquées. Le motif d'excitation correspondant au filtre générant cette impulsion est multiplié par l'amplitude de cette dernière, les autres impulsions sont alors comparées à une fraction r de ce motif calculé aux positions concernées. Plus précisément si on considère une impulsion « masquée » localisée à l'instant n_M et engendrée par le filtre k_M alors pour qu'une impulsion localisée à (n_P, k_P) soit mise à zéro il faut que :

$$x(n_P, k_P) \times E_{k,p} < r \times x(n_M, k_M) \times E_{kM}(n_p - n_M, k_P) \tag{5.5}$$

Puisque l'étape du masquage engendre une perte d'énergie due à la mise à zéro de certaines impulsions, dans [Feldbauer, 2005], les auteurs proposent une méthode de correction adaptative basée sur l'estimation des enveloppes des motifs d'excitation suivant une approche d'analyse par synthèse.

Si $\hat{x}(n_0, k_0)$ représente l'amplitude de l'impulsion à corriger, sa valeur corrigée $\hat{y}(n_0, k_0)$ est donnée par :

$$\hat{y}(n_0, k_0) = \hat{x}(n_0, k_0) \frac{\mathcal{E}(\sum_{l=0}^{N} x(n, l) * g_l(n) * h_k(n))(n_0, k_0)}{\mathcal{E}(\sum_{l=0}^{N} \hat{x}(n, l) * g_l(n) * h_k(n))(n_0, k_0)} \tag{5.6}$$

C'est à dire, que le motif d'excitation original ainsi que celui réduit sont passés à travers les filtres de synthèse avant d'en extraire les enveloppes en utilisant la transformée de Hilbert. Il faut noter que cette opération est effectuée par filtre d'analyse k et donc N fois pour un banc à N filtres. Les auteurs dans [Feldbauer, 2005] prétendent qu'il est possible de réduire le nombre d'impulsions pour un signal échantillonné à une fréquence de 16kHz par un facteur de 2/3 sans engendrer une dégradation de qualité notable. Cette méthode prend en considération le masquage post-stimuli et pré-stimuli (*backward masking*) ainsi que le masquage simultané ce qui permet de réduire considérablement le nombre d'impulsions à la sortie de l'étape du masquage. Cependant, la complexité de cette approche est très élevée et requière le stockage des motifs d'excitation pré-calculés ainsi qu'un nombre d'opérations de comparaisons élémentaires élevé. En effet, pour un banc contenant N filtres, il faut stocker N^2 motifs et extraire N enveloppes pour la correction des amplitudes des impulsions masquées. Dans [Thiemann, 2011], l'auteur confirme ces constatations et reporte la complexité d'une telle approche.

On propose dans ce chapitre une méthode plus simple qui ne nécessite pas le pré-calcul des motifs d'excitation et qui permet en exploitant aussi bien le masquage pré et post-stimuli que le masquage simultané pour réduire le nombre d'impulsions.

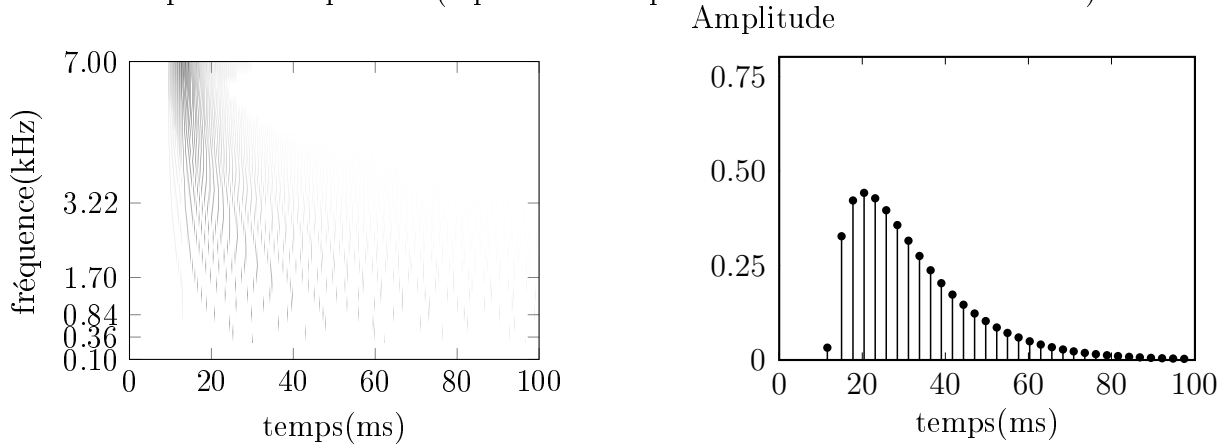
5.2 Nouveau modèle simple de masquage simultané

Cette section a pour but d'introduire un nouvel algorithme de masquage exploitant la redondance contenue dans les motifs d'excitation auditive. Pour ce faire, on commence par analyser le cas simple d'une impulsion de Dirac à l'entrée du banc de filtres ensuite on traite le cas d'un signal quelconque. On montre que cette approche, outre sa complexité computationnelle réduite, permet de réduire considérablement le nombre d'impulsions sans introduire une dégradation « agaçante » validée par l'algorithme PEAQ.

5.2.1 Le cas d'une impulsion de Dirac

La figure 5.1(b) présente le motif d'excitation auditive créé par une impulsion de Dirac. Dans cette expérience, la fréquence d'échantillonnage est de 16kHz et le nombre des filtres auditifs est de 32.

Comme tel qu'il est illustré sur la figure 5.1 cette représentation est redondante même pour un signal aussi simple qu'une impulsion de Dirac. Il est bien connu, qu'il existe une incertitude corrompant la mesure dans deux domaines conjoints particulièrement le domaine temporel et fréquentiel (représentés ici par les sorties du banc de filtres).



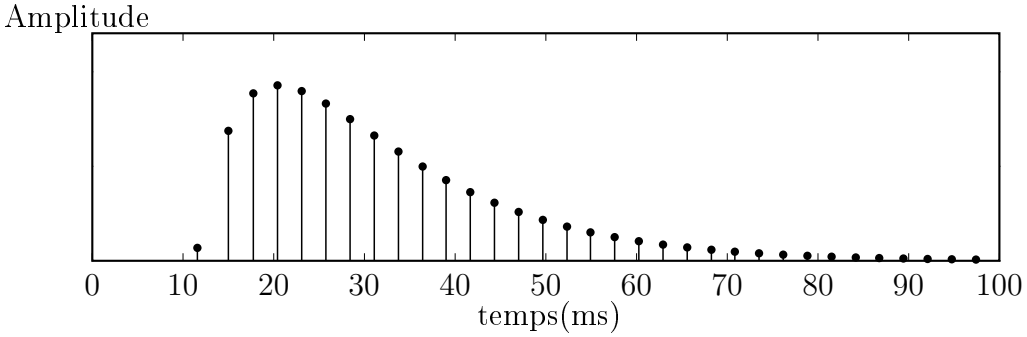
(a) Motif d'excitation d'une impulsion de dirac.

(b) Réponse d'un filtre auditif (F_c = 500Hz) suivi d'un modèle neuronal à une impulsion de dirac.

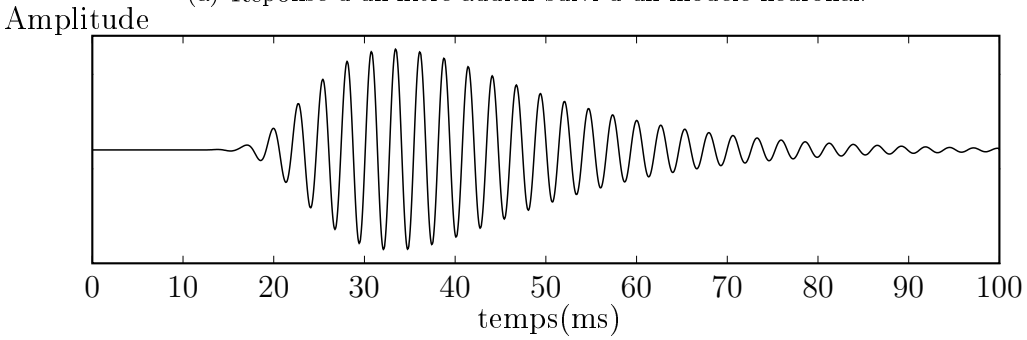
Figure 5.1 Motif d'excitation créé par une impulsion de dirac. Banc de 32 filtres avec une fréquence d'échantillonnage de 16kHz.

Cependant, la résolution temporelle du système présenté est bien plus élevée que sa résolution fréquentielle. En effet, l'étalement temporel des impulsions est bien évident dans la figure 5.1(b) où la réponse d'un filtre auditif suivi d'un modèle neuronal est donnée. La question qui se pose alors, jusqu'à où peut on réduire le nombre d'impulsions tout en étant capable de reconstruire l'impulsion de Dirac ? Pour répondre à cette question, rappelons

l'étape de synthèse. La sortie de chaque filtre du banc de filtres d'analyse une fois passée à travers le modèle neuronal est filtrée par le filtre de synthèse correspondant. Donc une condition nécessaire pour recréer le signal d'entrée à partir de sa représentation auditive réduite (après avoir mis certaines impulsions à zéro) consiste à être capable, pour chaque filtre, de récréer la même forme qu'aurait engendré le motif d'excitation non-réduit. La figure 5.2, représente la réponse impulsionnelle [1] du système proposé pour le filtre centré autour de 500Hz.



(a) Réponse d'un filtre auditif suivi d'un modèle neuronal.



(b) Réponse du filtre de synthèse pour une impulsion de dirac à l'entrée du banc de filtres.

Figure 5.2 Example d'analyse synthèse pour une excitation de dirac. La fréquence d'échantillonnage est de 16 kHz, la fréquence centrale du filtre auditif est de 500Hz.

Deux constations importantes sont à faire :

- La réponse du filtre de synthèse n'atteint pas sa valeur maximale à l'instant correspondant au maxima du motif d'excitation auditive.
- Cette réponse est encore plus étalée dans le temps que celle de la réponse impulsionnelle du filtre d'analyse.

1. Le système est non linéaire mais on peut considérer cette réponse comme étant celle qui approche la composante linéaire du système proposé.

Dans [Feldbauer, 2005] un modèle basé sur le motif d'excitation d'impulsion isolée et dans [Kubin et Kleijn, 1999b] un modèle de masquage exponentiel dont les paramètres empiriquement déterminés ont été utilisés pour réduire le nombre d'impulsions. On propose une méthode originale qui combine ces deux approches. On propose un seuil de masquage plus adapté à l'enveloppe temporelle des RIs des filtres de synthèse qu'il est possible d'estimer de manière récursive ce qui élimine la nécessité de stocker des motifs d'excitation pré-calculés et ne nécessite pas une approche itérative comme proposé dans [Feldbauer, 2005] ou dans [Thiemann, 2011].

On rappelle que l'expression de l'enveloppe des RIs des filtres de synthèse est donnée par :

$$\mathcal{E}(H) = A\exp(-\lambda kt)[1-\exp(-\lambda t)]^n \tag{5.7}$$

Où les valeurs numériques de λ et n ont été déterminées dans le chapitre 3.

5.2.2 Masquage post-stimuli

Se référant à l'équation (5.7), il est possible d'estimer l'enveloppe de n'importe quel train d'impulsions donné sans avoir à stocker en mémoire des motifs d'excitation unitaire comme dans [Feldbauer, 2005] ce qui se traduit par une réduction importante de la mémoire utilisée ainsi que des ressources computationnelles. On rappelle que la valeur maximale atteinte par l'enveloppe d'un BIT_n est donnée par :

$$t_{max} = log(m/k)/\lambda \tag{5.8}$$

Si on suppose une impulsion à l'entrée du filtre auditif se produisant à l'instant t, la valeur de l'enveloppe de l'excitation aux instants t et $t+1$ est donnée par :

$$\mathcal{E}(H(t)) = \delta_{t-t_{max}} A\exp(-\lambda k(t))[1-\exp(-\lambda(t))]^n \tag{5.9}$$

$$\mathcal{E}(H(t+1)) = \delta_{t-t_{max}} A\exp(-\lambda k(t+1))[1-\exp(-\lambda(t+1))]^n \tag{5.10}$$

C'est à dire pour une impulsion isolée se produisant à l'instant t, l'enveloppe de l'excitation atteint sa valeur maximale à l'instant $t+t_{max}$. En utilisant les équations précédentes il est possible de calculer l'enveloppe de l'excitation récursivement. En effet :

$$\mathcal{E}(H(t+1)) = \mathcal{E}(H(t))\exp(-\lambda)\left(\frac{1-\exp(-\lambda(t+1))}{1-\exp(-\lambda t)}\right)^n \tag{5.11}$$

Définissant r comme étant le facteur contrôlant d'étalement du seuil du masquage post-stimuli et en utilisant l'équation (5.11), il est possible de déterminer le seuil du masquage post-stimuli utilisant l'algorithme 5.1.

$$\lambda' = r\lambda \tag{5.12}$$

$$\mathcal{S}_{n+1}^{\text{post}} = \mathcal{S}_{n}^{\text{post}} \exp(-\lambda) \left(\frac{1 - \exp(-\lambda'(t+1))}{1 - \exp(-\lambda' t)} \right)^{n} \tag{5.13}$$

Si la valeur de $r = 0$, le seuil du masquage se comporte comme celui décrit dans [Kubin et Kleijn, 1999b](le seuil du masquage décroît exponentiellement). Quand la valeur de $r = 1$, le seuil du masquage suit exactement l'enveloppe de l'excitation engendrée par le filtre d'analyse et donc constitue le motif d'excitation d'impulsion isolé proposé par [Feldbauer, 2005] (quand on ne considère pas l'interaction entre filtres adjacents).

L'algorithme 5.1 donne les étapes pour calculer de manière récursive le seuil du masquage post-stimuli. Si la variable y dénote le train d'impulsions du filtre auditif concerné, l'algorithme consiste à parcourir les valeurs des impulsions et à calculer récursivement le seuil du masquage. Un compteur est utilisé pour garder en mémoire le délai entre l'instant pendant lequel une impulsion a dépassé le seuil du masquage et l'instant courant. Si ce délai dépasse la valeur de t_{max} (équation (5.8)), la valeur du seuil du masquage est estimée selon l'équation (5.13). Une fois le seuil du masquage estimé, la valeur de l'impulsion est comparée à ce dernier. Si la valeur de l'impulsion dépasse le seuil du masquage, la valeur de ce dernier est remplacée par la valeur de l'impulsion et le compteur est remis a zéro. Dans le cas contraire, le compteur est augmenté et l'algorithme passe à l'impulsion suivante.

Le résultat de l'application de l'algorithme 5.1 aux impulsions données par la figure 5.2 est donné par la figure 5.3.

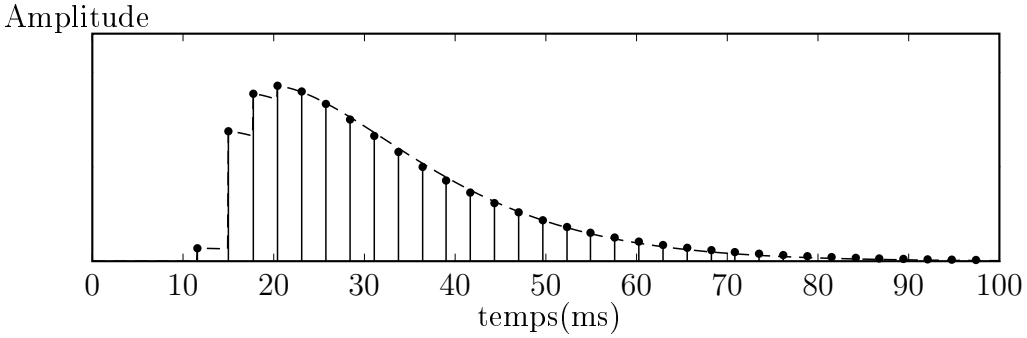


Figure 5.3 Seuil de masquage post-stimuli (ligne discontinue) estimé selon l'algorithme 5.1. La fréquence d'échantillonnage est de 16 kHz, la fréquence centrale du filtre auditif est de 500Hz.

Algorithme 5.1 : Algorithme d'estimation du seuil du masquage post-stimuli

Entrées :
y : Sortie d'un filtre auditif k passée à travers le modèle neuronal pour la trame courante.
$\mathcal{S}_{n-1}^{\text{post}'}$: Dernière valeur du seuil de masquage post-stimuli pour le filtre k pour la trame précédente.
Sorties :
$\mathcal{S}_{n}^{\text{post}}$: Seuil de masquage post-stimuli pour le filtre k pour la trame courante.
Données :
λ : Paramètre du filtre auditif k (équation (5.7)).
n_y : Durée de la trame d'analyse.
ctr : Compteur incrémenté à partir de l'instant où une impulsion dépasse le seuil de masquage dans la trame précédente.
r : Paramètre contrôlant l'étalement du seuil de masquage.
c : Compression logarithmique donnée dans l'équation (4.1)
On définit $\lambda' = -\lambda r$
pour $i = 1 : n_y$ **faire**
  si $ctr > \log(4)/\lambda'$ *et* $i > 1$ **alors**
    $t = \dfrac{ctr + \log(4)/\lambda'}{Fs}$
    $\beta = \exp(\lambda) \left(\dfrac{1 - \exp(\lambda' t)}{1 - \exp(\lambda'(t + 1/Fs))} \right)^n$
    $S = (\beta \times \mathcal{S}_{n}^{\text{post}}(i-1))^c$
  sinon
    $S = \mathcal{S}_{n-1}^{\text{post}'}$
  fin
  si $y(i) \geq S$ **alors**
    $\mathcal{S}_{n}^{\text{post}}(i) = y(i)$
    $ctr = 0$
  sinon
    $\mathcal{S}_{n}^{\text{post}}(i) = S$
    $ctr = ctr + 1$
  fin
fin

Si on considère le cas de la figure 5.3, n'importe quelle valeur de $r < 1$ permet de réduire considérablement le nombre d'impulsions à la sortie du bloc du masquage. Cependant, si on regarde le motif d'excitation synthétisé à partir des impulsions restantes, il est évident que la simple correction des amplitudes des impulsions n'est pas suffisante pour recréer la même excitation générée par le train d'impulsions complet.

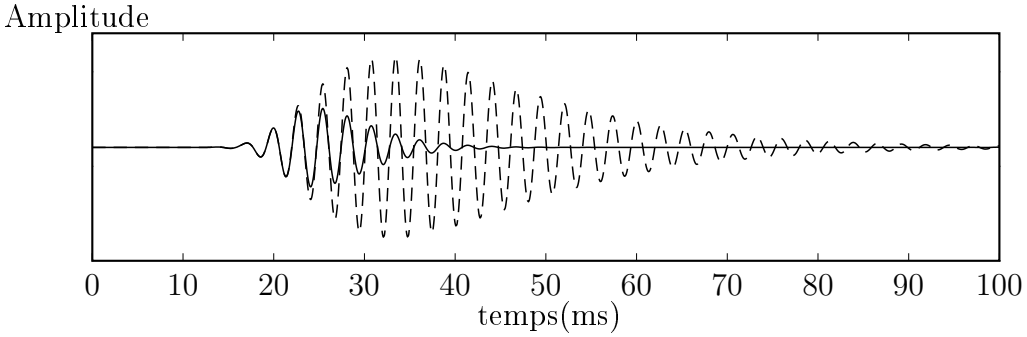


Figure 5.4 Excitations avant et après l'application du masquage post-stimuli. Le motif réduit est représenté par une ligne continue. La fréquence centrale du filtre auditif est de 500Hz.

En effet, les deux excitations (complète et réduite) n'atteignent pas leurs maxima aux même instants. Ceci est principalement dû aux impulsions survenant au début de ce motif d'excitation. La figure 5.4 montre clairement que les excitations coïncident au début de celles-ci. Le fait de changer les amplitudes des impulsions du motif réduit ne permettrait pas à la fois de « déplacer » l'instant auquel l'excitation réduite atteint son maximum et de conserver l'énergie perdue due à la mise à zéros des impulsions masquées. Pour remédier à ce problème on propose alors de mettre des impulsions au début de l'excitation à zéro dans le but d'introduire un « délai » artificiel ce qui a pour but de faire coïncider le maximum de l'excitation réduite avec celle de l'excitation complète. Dans la section suivante, on aborde le masquage pré-stimuli.

5.2.3 Masquage pré-stimuli

De la même manière, comme décrit dans la section 5.2.2, on peut estimer de façon récursive le seuil du masquage pré-stimuli. Il suffit dans ce cas de parcourir le train d'impulsions dans le sens inverse et d'estimer l'enveloppe de l'excitation. Utilisant les mêmes notations que dans l'équation (5.13), le seuil du masquage pré-stimuli est estimé par :

$$\mathcal{S}_{n+1}^{\text{pre}} = \mathcal{S}_{n}^{\text{pre}} \exp(-\lambda) \left(\frac{1 - \exp(-\lambda' t)}{1 - \exp(-\lambda'(t-1))} \right)^{n} \tag{5.14}$$

Algorithme 5.2 : Algorithme d'estimation du seuil du masquage pré-stimuli

Entrées :
y : Sortie d'un filtre auditif k passée à travers le modèle neuronal pour la trame courante.
Sorties :
$\mathcal{S}_n^{\text{pre}}$: Seuil de masquage pré-stimuli pour le filtre k pour la trame courante.
Données :
λ : Paramètre du filtre auditif k (équation (5.7)).
n_y : Durée de la trame d'analyse.
ctr : Compteur incrémenté à partir de l'instant où une impulsion dépasse le seuil de masquage.
r : Paramètre contrôlant l'étalement du seuil de masquage.
c : Compression logarithmique donnée dans l'équation (4.1)
On définit $\lambda' = -\lambda r$
$ctr = 0$
pour $i = n_y : 1$ **faire**
  $t = \max(0, \dfrac{-ctr + \log(4)/\lambda'}{Fs})$
  $\beta = \exp(\lambda) \left(\dfrac{1 - \exp(\lambda' t)}{1 - \exp(\lambda'(t - 1/Fs))} \right)^n$
  $S = (\beta \times \mathcal{S}_n^{\text{pre}}(i-1))^c$
  si $y(i) \geq S$ **alors**
    $\mathcal{S}_n^{\text{pre}}(i) = y(i)$
    $ctr = 0$
  sinon
    $\mathcal{S}_n^{\text{pre}}(i) = S$
    $ctr = ctr + 1$
  fin
fin

L'algorithme 5.2 décrit les étapes nécessaires pour estimer le seuil du masquage pré-stimuli. Notez la différence avec l'algorithme 5.1 : les impulsions sont parcourues dans le sens inverse et de ce fait la valeur de la variable *ctr* est soustraite à celle de t_{max} pour mimer le comportement d'un *look ahead buffer*. L'application de l'algorithme du masquage pré-stimuli au train d'impulsions de la figure 5.2(a) est donné sur la figure 5.5. Le masquage

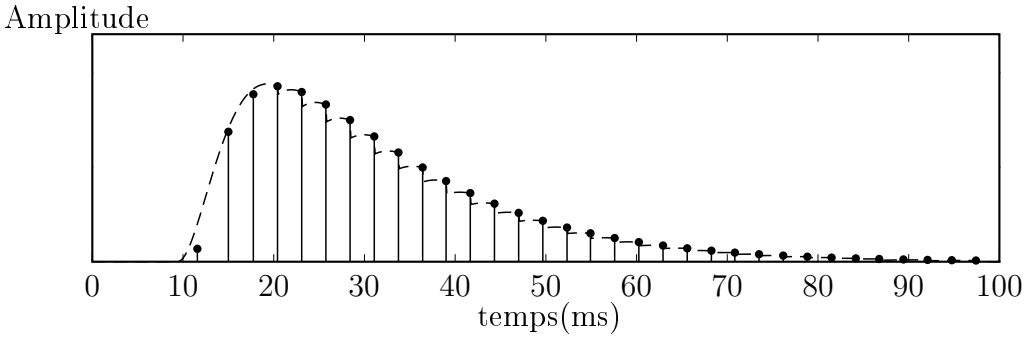


Figure 5.5 Seuil de masquage pré-stimuli (ligne discontinue) estimé selon l'algorithme 5.2. La fréquence d'échantillonnage est de 16 kHz, la fréquence centrale du filtre auditif est de 500Hz.

temporel $\mathcal{S}^{\text{tem}}$ peut être alors estimé comme étant :

$$\mathcal{S}^{\text{tem}} = \max(\mathcal{S}^{\text{pre}}, \mathcal{S}^{\text{post}}) \tag{5.15}$$

5.2.4 Masquage simultané

Le masquage simultané se produit lorsqu'un son est rendu inaudible par un autre joué simultanément. L'effet engendré par le deuxième son masquant dépend de ses caractéristiques fréquentielles et temporelles. Par example, dans les expériences du masquage par un bruit blanc à bandes étroites non seulement le niveau du bruit masquant importe mais aussi son étendue fréquentielle (voir chapitre 3). Il est bien connu que la courbe du masquage simultané est non symétrique et sa forme dépend de l'intensité du signal masquant [Glasberg, 2002; Moore, 2012]. Le masquage simultané a été utilisé par example dans nombreux codecs comme le codeur MP-3 et MP-4 [Brandenburg et Bosi, 1997; Wolters *et coll.*, 2003]. Alors que dans les codecs cités plus hauts, l'estimation du seuil du masquage se fait dans le domaine fréquentiel, dans le travail présenté le masquage simultané est estimé à partir des trains d'impulsions. Si $\mathcal{S}_{ch}$ représente le seuil du masquage estimé à la sortie du filtre auditif *ch*, le seuil du masquage simultané est estimé selon :

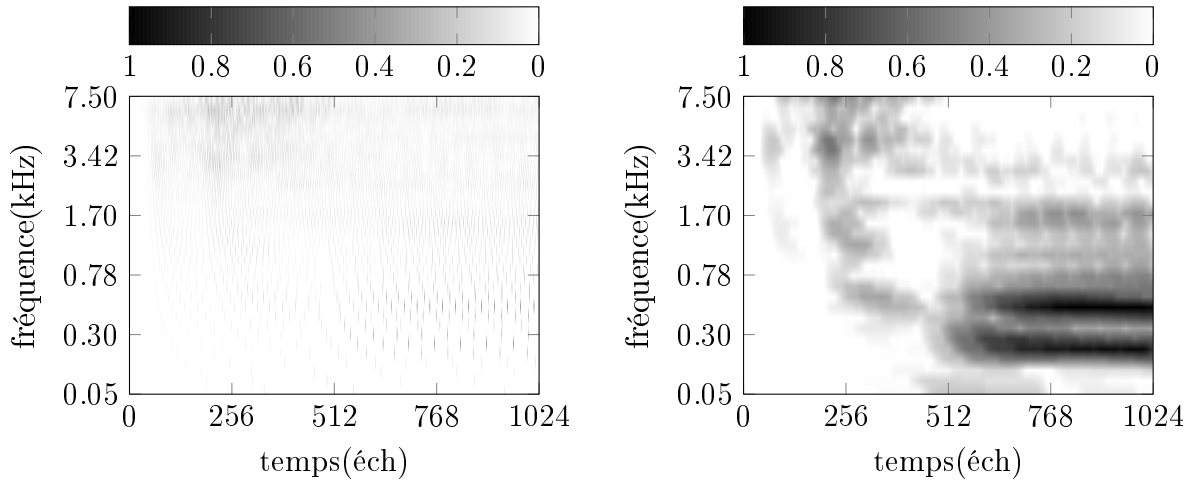
$$\mathcal{S}^{\text{sim}}_{ch} = \max(\exp^{-c0(f_{ch+1}-f_{ch})}\,\mathcal{S}_{ch+1}, \exp^{-c1(f_{ch}-f_{ch-1})}\,\mathcal{S}_{ch-1}) \tag{5.16}$$

Où f_{ch} représente la fréquence centrale du filtre auditif concerné et $c0$ et $c1$ des constantes décrivant la pente de décroissance de ce seuil. Le seuil du masquage simultané décrit par l'équation (5.16) suppose que le seuil du masquage fréquentiel décroît exponentiellement. Il est à noter que cette hypothèse est très simplificatrice, mais comme il a été reporté dans [Glasberg, 2002; Kubin et Kleijn, 1999b; Moore, 2012], le gain dû au masquage simultané est minimal comparativement au masquage temporel, le modèle décrit par (5.16) semble être suffisant.

Une fois les seuils du masquage temporel et simultané estimés, il est possible d'estimer le seuil du masquage global $\mathcal{S}$:

$$\mathcal{S} = \max(\mathcal{S}^{\text{tem}}, \mathcal{S}^{\text{sim}}) \tag{5.17}$$

La figure 5.6 présente un example de l'application des équations citées plus haut pour



(a) Train d'impulsions extraits d'une trame de signal de parole.

(b) Seuil de masquage temporel et simultané estimé à partir du train d'impulsions de la figure 5.6(a).

Figure 5.6 Estimation du seuil de masquage temporel et simultané à partir d'une trame de signal de parole.

estimer le seuil du masquage à partir du train d'impulsions de la figure 5.6(a).

Le seuil du masquage $\mathcal{S}$ une fois estimé selon l'équation (5.17), peut être utilisé pour classifier les impulsions en deux classes : impulsions masquantes et impulsions masquées notées respectivement y_1 et y_0 et données par :

$$y(n) \in \begin{cases} y_1 \text{ si } y(n) \geq \mathcal{S}_n \\ y_0 \text{ si } 0 < y(n) < \mathcal{S}_n \end{cases} \tag{5.18}$$

Les paramètres r, $c0$ et $c1$ dans les équations (5.13) et (5.16) permettent de contrôler l'étalement du seuil du masquage et donc par la même occasion permettent de contrôler la partition des valeurs des impulsions entre impulsions masquantes et impulsions masquées. Si la valeur des amplitudes des impulsions masquées est mise à zéro, une estimation du taux de parcimonie τ_s est donnée par l'équation (5.19).

$$\tau_s = \lim_{n_y \to \infty} \frac{|y_1|}{|y_0| + |y_1|} \tag{5.19}$$

Où $|.|$ dénote l'opération qui donne le nombre d'éléments dans un ensemble donné et n_y dénote la longueur de la trame analysée. Le nombre d'impulsions par échantillon peut être aussi estimé selon l'équation (5.20).

$$\tau_{\text{éch}} = \lim_{n_y \to \infty} \frac{|y_1|}{n_y} \tag{5.20}$$

Cependant, la mise à zéro de certaines impulsions engendre nécessairement une perte d'énergie. Dans la section suivante, on propose une méthode pour compenser cette perte.

5.3 Correction adaptative des amplitudes des impulsions masquantes

Le bloc du masquage engendre une perte d'énergie. Dans le but de synthétiser le signal à partir de sa représentation éparse, il est nécessaire de modifier l'amplitude des impulsions restantes dans le but de restaurer la distribution d'énergie dans le domaine temps-fréquence. Dans cette section, on propose une nouvelle approche pour compenser cette perte d'énergie. Alors que dans [Feldbauer, 2005] une méthode computationnellement intensive basée sur l'estimation des motifs d'excitation est utilisée (voir section 5.1), on utilise une approche plus simple inspirée de la méthode de correction donnée par l'équation (4.7).

Si $y_1(i)$ dénote la valeur de l'amplitude d'une impulsion masquante, sa valeur $\hat{y}_1(i)$ modifiée est donnée par :

$$\hat{y}_1(i) = y_1(i) + \sum_{k=0}^{m} \frac{fs/f_c}{|t_i - t_k|} y_0(k) \ \forall k \in \Omega_i \tag{5.21}$$

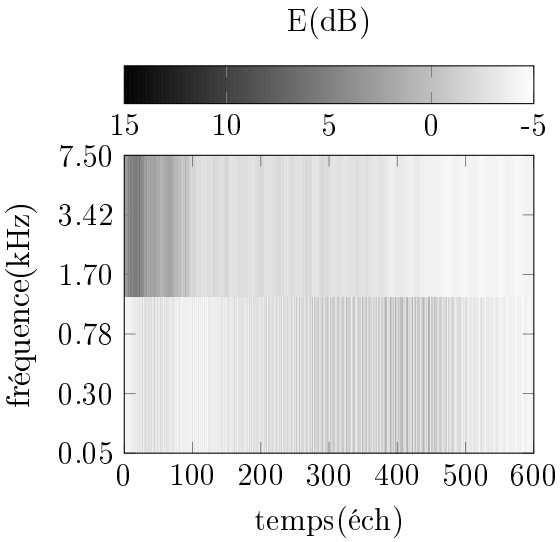
Où f_s et f_c représentent respectivement la fréquence d'échantillonnage et la fréquence centrale du filtre auditif concerné. Ω_i représente l'ensemble des impulsions masquées par l'impulsion $y_1(i)$. La différence $t_i - t_k$ exprimée en échantillons représente la durée séparant

l'impulsion masquante $y_1(i)$ de l'impulsion masquée $y_0(k)$. Cette durée ne peut être nulle puisqu'une impulsion ne peut être à la fois masquante et masquée.
L'équation (5.21) permet de modifier de façon adaptative la valeur de l'amplitude de l'impulsion masquante en y additionnant une somme pondérée des impulsions masquées. Le terme pondérant est proportionnel à la durée séparant impulsions masquantes et impulsions masquées. Intuitivement, plus cette durée est élevée plus la contribution de l'impulsion masquée est minimale. Cette opération est appliquée à toutes les impulsions masquantes et permet de façon adaptative de restaurer l'énergie perdue suite à l'élimination des impulsions masquées.

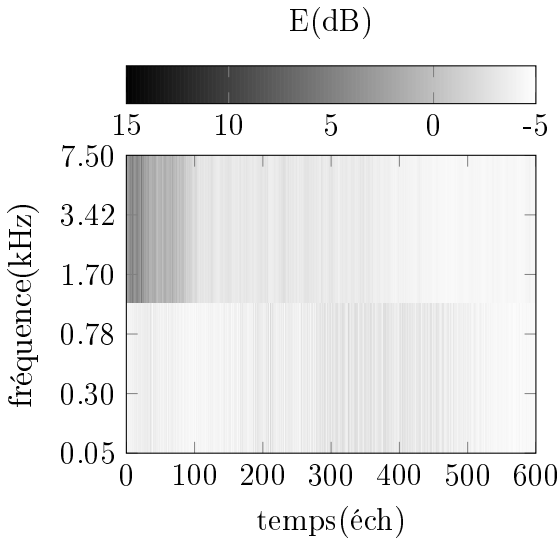
La figure 5.7 donne un example de l'application de cette méthode simple de correction d'amplitude. Sur les figures 5.7(b) et 5.7(a) la différence entre excitations à la sortie du banc de filtres de synthèse est donnée sur une échelle logarithmique. La figure 5.7(b) représente cette différence quand la méthode de correction d'amplitude (équation (5.21)) est appliquée. L'effet de cette correction est bien visible autour de la fréquence 1kHz où la différence entre excitations est bien plus petite que celle sur la figure 5.7(a) où la méthode de correction n'a pas été appliquée. L'application de cette méthode de correction sur le tain d'impulsions à la sortie du filtre auditif centré autour de 1kHz est illustrée sur la figure 5.7(c).

5.4 Nouvelle structure du codec proposé

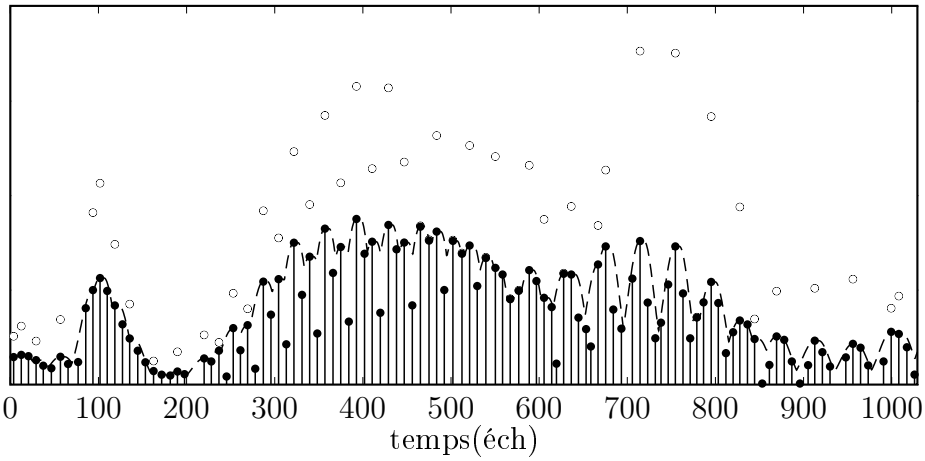
Dans les sections précédentes on a présenté les algorithmes qui permettent de réduire le nombre d'impulsions à la sortie du modèle neuronal. On a aussi présenté une méthode simple pour la correction des amplitudes des impulsions masquantes dans le but de restaurer l'énergie perdue suite à l'élimination des impulsions masquées. La nouvelle structure du codec proposé est donnée sur la figure 5.8. Le signal à l'entrée est analysé sous forme de trames. Chaque trame est passée à travers le banc de filtres d'analyse ensuite à travers le modèle neuronal. Le modèle neuronal dont le fonctionnement est décrit par l'équation (4.1) agit comme un sous-échantillonneur adaptatif. Une fois les trains d'impulsions extraits, les seuils du masquage sont déterminés selon les équations (5.13) et (5.14). Ces seuils sont utilisés par la suite pour classer les impulsions en impulsions masquantes et masquées. Alors que les valeurs des impulsions masquées sont mises à zéro, celles des impulsions masquantes sont amplifiées pour restaurer la perte d'énergie (équation (5.21)). Les impulsions survivantes par la suite sont passées à travers le filtre d'égalisation qui n'est qu'un simple gain/délai par filtre. Les paramètres du filtre égaliseur sont déterminés selon les équations (4.12) et (4.18) pour un délai et un nombre de filtres donné. La dernière



(a) Différence d'excitations entre trains d'impulsions avant et après l'application du seuil de masquage.



(b) Différence d'excitations entre trains d'impulsions avant et après l'application du seuil de masquage en intégrant la correction des amplitudes.



(c) Example de correction d'amplitudes des impulsions masquantes en utilisant l'équation (5.21).

Figure 5.7 Différence d'excitations entre train d'impulsions réduit et corrigé. Dans la figure 5.7(c), le seuil de masquage est représenté par une ligne discontinue alors que l'amplitude des impulsions après correction utilisant l'équation (5.21) est représentée par le symbole o. La fréquence du filtre auditif est de 1 kHz.

étape de synthèse consiste à passer les impulsions égalisées à travers le banc de filtres de synthèse. Le signal synthétisé est récupéré en sommant les sorties du banc de filtres de synthèse.

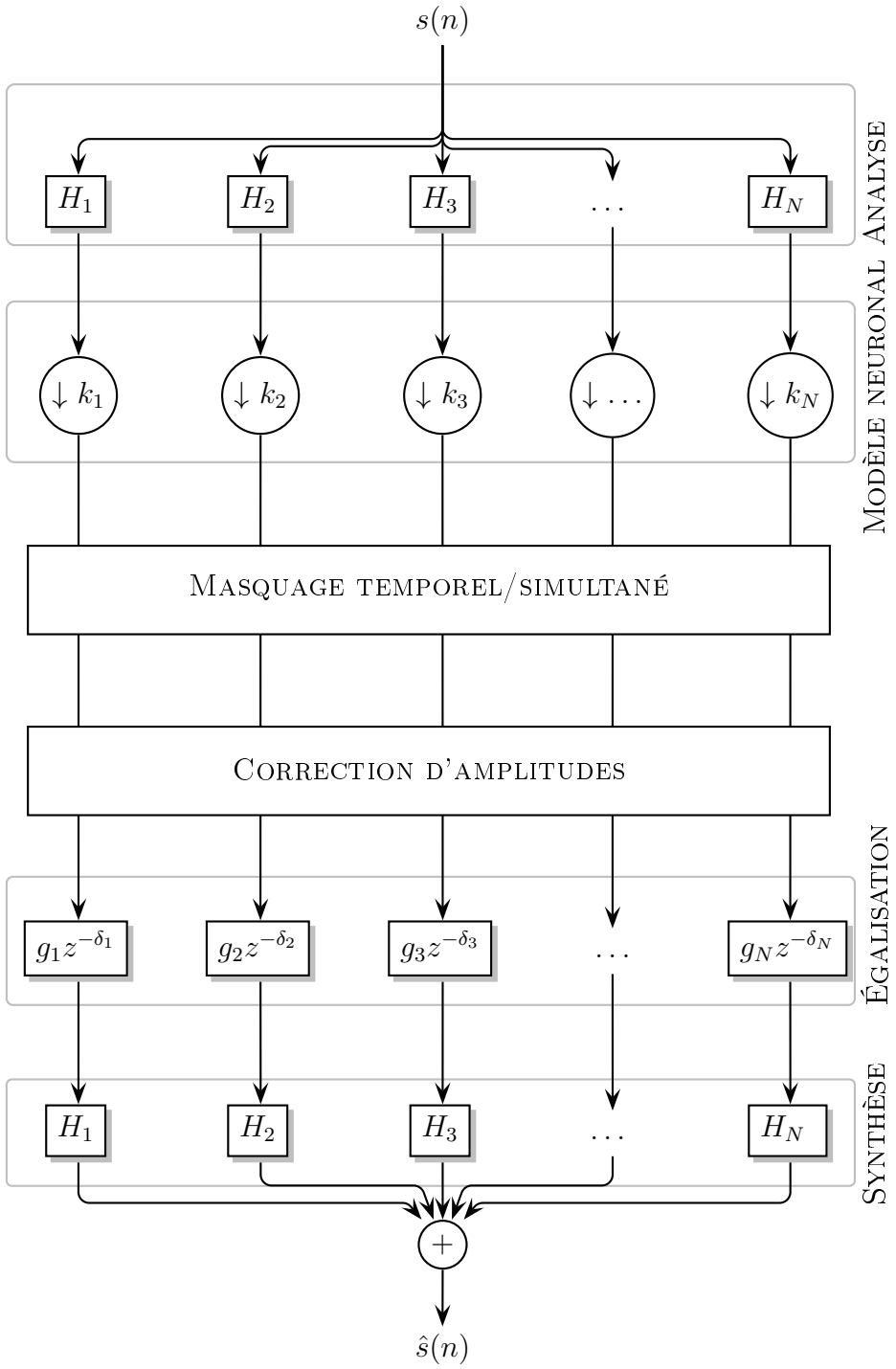


Figure 5.8 Structure du codec proposé incluant les modèles neuronaux et la correction d'amplitudes adaptative.

5.5 Résultats expérimentaux

Dans cette section, on propose d'estimer la qualité du système proposé en utilisant l'algorithme PEAQ (section 4.4). Les signaux de parole utilisés dans cette section proviennent de la collection de signaux de parole collectés par Texas Instruments et Massachusetts Institute of Technology (TIMIT)[2]. TIMIT contient des enregistrements large bande de 630 locuteurs de huit dialectes majeurs de l'anglais américain, chacun lisant dix phrases phonétiquement riches. Ce corpus comprend aussi les transcriptions orthographiques et phonétiques de chaque fichier. Les fichiers audio ont une résolution de 16 bits par échantillon et sont échantillonnés à une fréquence de 16kHz. Les sous-ensembles d'entraînement et de test, équilibrés pour la couverture phonétique et dialectale, sont aussi spécifiés [Garofolo *et coll.*, 1993].

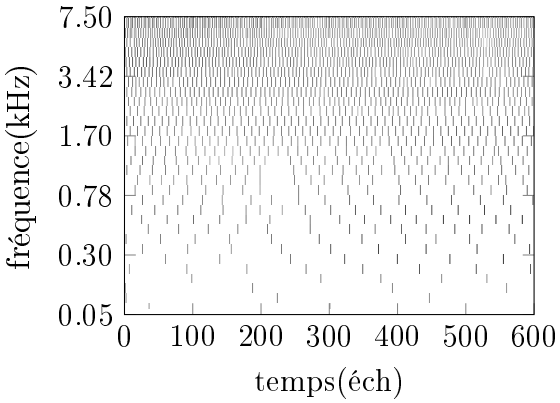
Dans le but d'aboutir à des résultats concluants, les expériences doivent être reproductibles et généralisables. Aussi bien dans [Kubin et Kleijn, 1999b] que dans [Feldbauer, 2005], les fichiers audio utilisés pour caractériser leurs systèmes respectifs ne sont ni donnés ni spécifiés. Une vague description est donnée dans [Feldbauer, 2005] (section 3.5.2.2) : « *For the speech material 12 English sentences spoken by six native male and six native female speakers were chosen. The material had been recorded at a sampling rate of 16 kHz and with 16 bits accuracy. The duration of these speech samples varied from 2.2 to 4 seconds.* ». Ceci rend la comparaison avec leur approche difficile voire impossible. Outre le fait que 36 secondes (3 secondes par signal) de signaux de parole n'est pas suffisante pour caractériser un système ni estimer ses performances, l'estimation des taux de compression $\tau_{\text{éch}}$ et τ_s dependent fortement de la quantité de parole nette (excluant les périodes de silence). Il suffit d'imaginer le cas extrême d'un signal nul, dans ce cas par example le nombre d'impulsions par échantillon est zéro ! Dans [Thiemann, 2011] reporte que la performance de leur système dépend des signaux analysés (voir la discussion au chapitre 4).

On propose dans les sections suivantes de sous-échantillonner TIMIT pour en construire deux ensembles. Un ensemble pour déterminer les paramètres optimaux du système proposé, un autre ensemble pour valider ces paramètres. Pour ce faire, quatre-vingt locuteurs aléatoirement choisis couvrant les huit dialectes disponibles sont partagés entre les deux ensembles. Chaque ensemble contient vingt hommes et vingt femmes lisant un text différent à chaque fois. La durée totale de ces signaux est de 5 minutes. Ces signaux étant ainsi choisis permettent de donner une estimation fiable et non biaisée de la qualité du système proposé. Une description détaillée de ces signaux est donnée en annexe A.

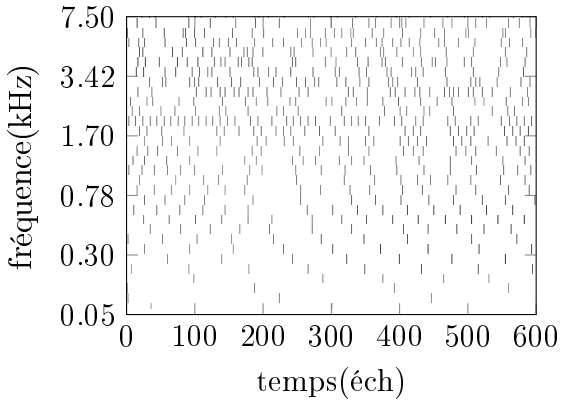
2. `https://catalog.ldc.upenn.edu/LDC93S1`

5.5.1 Qualité du nouveau modèle du masquage

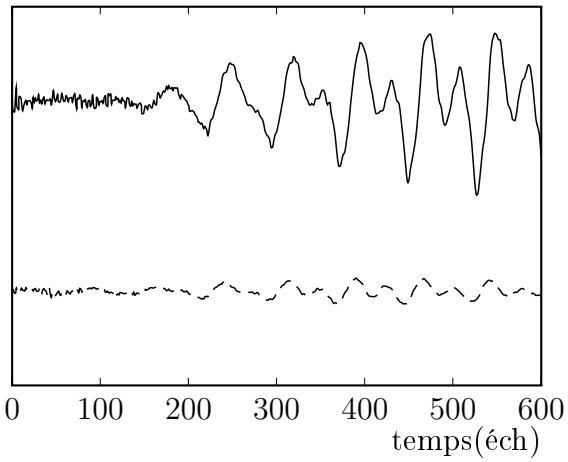
La figure 5.9 donne un example de synthèse d'un signal de parole à partir de son motif d'excitation auditive complet et réduit. Les figures 5.9(c) et 5.9(d) donnent l'erreur de reconstruction calculée comme étant la différence avec le signal original. Même si le RSB est relativement moyen (14dB) l'erreur engendrée par la reconstruction à partir du motif d'excitation réduit est inaudible. Dans l'example donné dans la figure 5.9, la valeur du



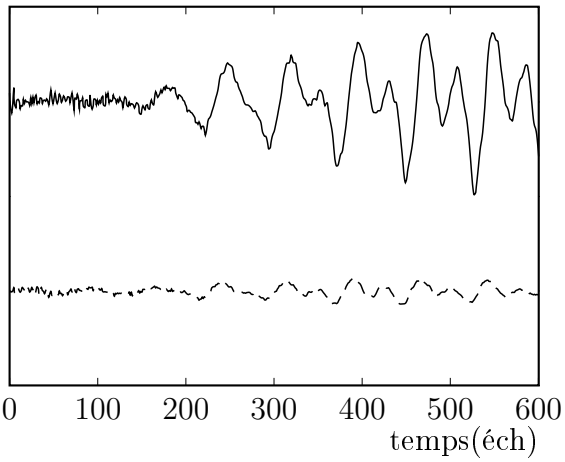
(a) Motif d'excitation complet d'une trame de signal parole.



(b) Motif d'excitation réduit d'une trame de signal parole.



(c) Example de synthèse d'une trame de signal de parole à partir de son motif d'excitation complet.



(d) Example de synthèse d'une trame de signal de parole à partir de son motif d'excitation réduit.

Figure 5.9 Example de synthèse de signaux à partir de leurs motifs d'excitation auditive complets et réduits. Dans chaque cas, l'erreur calculée comme différence par rapport au signal d'origine est donnée en ligne discontinue. Le nombre des filtres auditifs est de 32. Dans le cas présenté dans cette figure le nombre d'impulsions est réduit par un facteur de 70%.

paramètre r contrôlant le seuil du masquage est de 0.7. Avec cette valeur, le taux de parci-

monie τ_s vaut 70%. Le paramètre r dans l'équation (5.14) contrôlant l'étalement temporal du masquage pré-stimuli ainsi que le paramètre r dans l'équation (5.13) contrôlant l'étalement du seuil du masquage post-stimuli sont variés indépendamment et sont notés r_{pre} et r_{post} respectivement. La séparation entre les deux paramètres permet de vérifier l'effet de chaque paramètre sur la qualité de synthèse indépendamment.

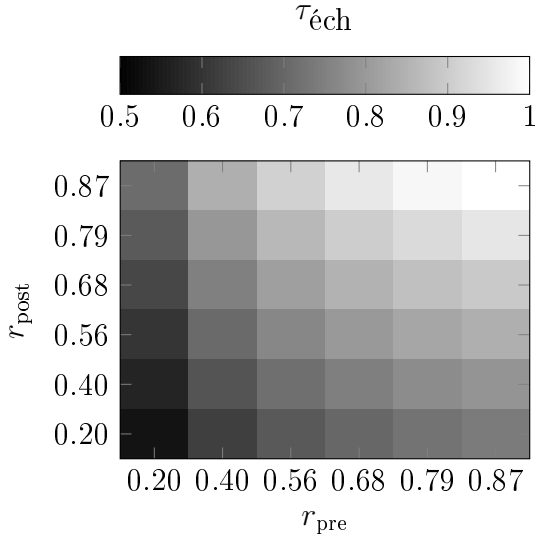
Pour des valeurs données r_{pre} et r_{post}, les signaux de l'ensemble d'entraînement (voir annexe A.1) sont passés à travers le système proposé. Les seuils du masquage sont évalués utilisant les valeurs de r_{pre} et r_{post} pour réduire le nombre d'impulsions ensuite les signaux sont synthétisés à partir des impulsions restantes (voir le diagramme 5.8). Cette opération est effectuée pour chacun des quarante signaux de l'ensemble d'entraînement. La valeur ODG de chaque signal est ensuite estimée utilisant l'algorithme PEAQ. Pour cette expérience le nombre de filtres auditifs est de 32, le délai du système est de 25ms et la taille de la trame d'analyse est de 25ms (400 échantillons pour une fréquence d'échantillonnage de 16kHz).

La figure 5.10 donne les résultats de cette expérience. La figure 5.10(a) donne le nombre moyen d'impulsions par échantillon et la figure 5.10(b) donne la valeur de l'ODG correspondant. On rappelle que cette dernière métrique donne une estimation de la note que donnerait un auditeur en comparant un signal référence à un signal test. Le tableau 5.1 donne l'interprétation des valeurs de l'ODG telle que décrite dans [ITU-R BS. 1284-1, 2002].

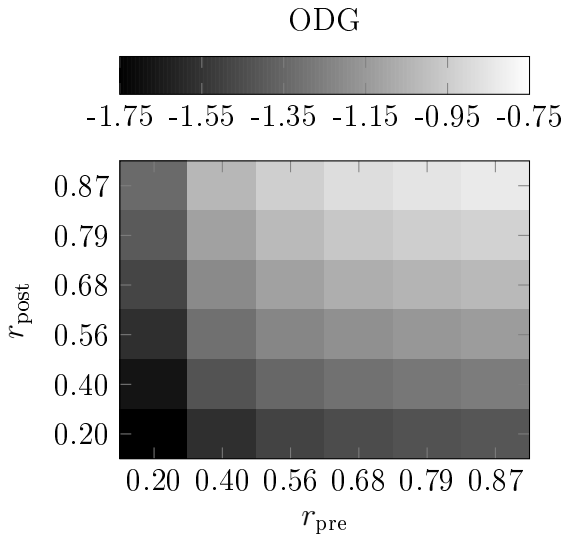
Tableau 5.1 Interprétation des valeurs de l'ODG.

Rec ITU-R BS. 1284-1	ODG	Dégradation	Qualité
5.0	0.0	Imperceptible	Excellente
4.0	-1.0	Perceptible non-agaçante	Bonne
3.0	-2.0	Légèrement agaçante	Moyenne
2.0	-3.0	Agaçante	Mauvaise
1.0	-4.0	Très agaçante	Médiocre

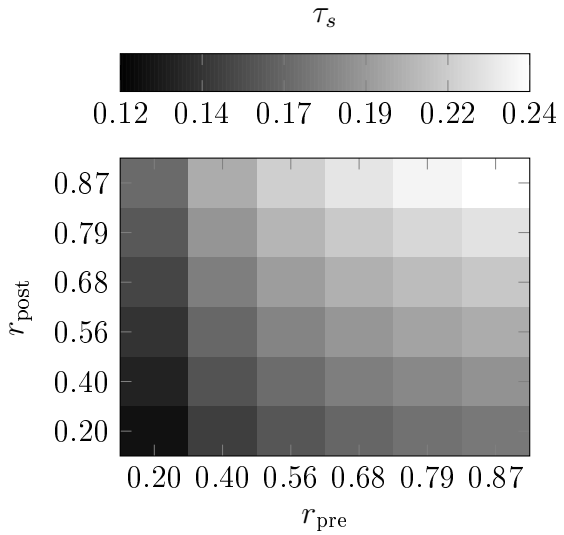
Pour une valeur de $\tau_{\text{éch}}$ valant 1 et correspondant à $r_{\text{pre}} = r_{\text{post}} = 0.87$, l'ODG moyen vaut -0.75 signifiant une qualité de reconstruction bonne voire excellente. Pour une valeur de r_{pre} fixée, diminuer la valeur de r_{post} permet de réduire la valeur de $\tau_{\text{éch}}$. Cela aussi est valide quand la valeur de r_{post} est fixée et la valeur de r_{post} est diminuée. Il semble qu'il n'y a pas de préférence quand il s'agit de décider quel paramètre permet de réduire la valeur de $\tau_{\text{éch}}$ sans introduire une dégradation et les deux paramètres r_{pre} et r_{post} semblent affecter la qualité de synthèse linéairement. Quand la valeur de $r_{\text{pre}} = r_{\text{post}} = 0.2$ le nombre moyen d'impulsions par échantillon est de 0.5 et la qualité de synthèse est moyenne. Pour



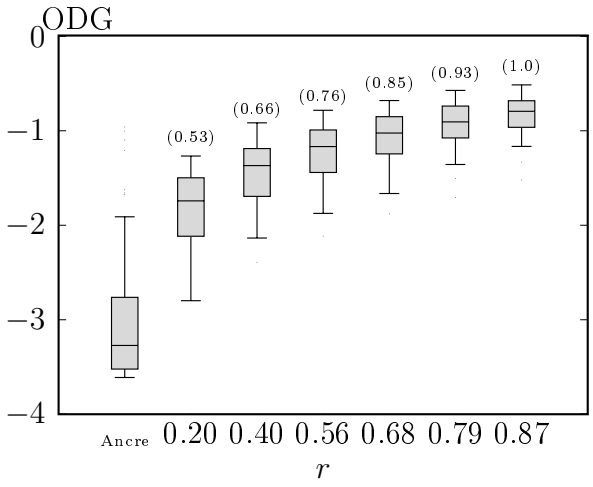
(a) Nombre moyen d'impulsions par échantillon pour différents paramètres du seuil de masquage.



(b) ODG moyen pour différents paramètres du seuil de masquage.



(c) Taux de parcimonie τ_s pour différents paramètres du seuil de masquage.



(d) ODG moyen pour différents paramètres du seuil de masquage.

Figure 5.10 Nombre d'impulsions par échantillon et ODG moyen pour différents paramètres du seuil de masquage. Sur la figure 5.10(d) le nombre moyen d'impulsions par échantillon $\tau_{\text{éch}}$ est donné entre parenthèses.

cette même valeur le taux de parcimonie est de 12% : Quasiment 88% des impulsions sont éliminées et la qualité de synthèse reste moyenne.

La figure 5.10(d) présente l'ODG moyen des signaux de validation quand les valeurs de r_{pre} et r_{post} sont variées par la même quantité. Le nombre moyen d'impulsions par échantillon est donné sur la même figure. Sur la même figure, l'ODG moyen du signal ancre est donné. On rappelle que le signal ancre est tout simplement une version filtrée du signal original. Dans cette expérience, le signal ancre est créé en filtrant le signal original par un filtre passe-bas limité à 4kHz mimant ainsi un test MUSHRA. Contrairement aux résultats reportés par [Thiemann, 2011] la qualité de synthèse du système proposé est toujours meilleure que celle obtenue par le signal filtré. Même pour un nombre d'impulsions par échantillon aussi petit que 0.53 la qualité de synthèse est statistiquement meilleure. Le diagramme en boîte (boîte de Tukey) décrit la dispersion statistique des résultats obtenus. Le rectangle sur la figure joint le premier au troisième quartile et est coupé par la médiane. Le premier et le neuvième décile (D1/D9) sont aussi donnés et sont représentés par des segments de droite. La figure 5.10(d) montre aussi que le paramètre r permet de contrôler linéairement la qualité de synthèse en contrôlant le nombre d'impulsions par échantillon. Pour une valeur de $r = 0.56$, l'ODG moyen vaut -1.25 impliquant une qualité de synthèse généralement bonne. Comme expliqué plus haut, il n'est pas possible de comparer les résultats obtenus dans ce travail avec ceux reportés dans [Kubin et Kleijn, 1999b] ni avec ceux de [Feldbauer, 2005]. Mais si on suppose que dans ces travaux les même signaux sont utilisés pour estimer la qualité des systèmes proposés, le tableau suivant donne une « comparaison » approximative. Dans [Feldbauer, 2005] seule la valeur de 0.9 impulsions par échantillon a été validée par des tests d'écoute. La valeur de 0.66 citée dans le même ouvrage quant à elle n'est validée que par un test d'écoute informel : « *We performed experiments with narrowband-filltered speech signals and the 20-channel filterbank setup with center frequencies from 100 Hz to 3600 Hz...Depending on the chosen input speech sample, the benefit of incorporating the adaptation circuit is a reduction in the number of pulses between 15% and almost 21% while producing the same perceptual quality after the resynthesis as confirmed by an informal listening test.* »

La figure 5.11 présente les résulats de l'ensemble de validation. Les résultats sont séparés par sexe. Pour n'importe quelle valeur de r, la distribution des valeurs de l'ODG entre les deux catégories est bien visible : les signaux de parole issus d'interlocuteurs masculins ont généralement en moyenne des ODGs supérieurs à ceux issus d'interlocuteurs féminins. Cette constatation est aussi reportée dans [Feldbauer, 2005]. Pour des valeurs de r proches de 1, la différence entre les deux groupes vaut -0.25. Quand la valeur de r est proche de 0,

Tableau 5.2 Comparaison entre différents systèmes de synthèse de signaux de parole à partir de leurs motifs d'excitation auditive.

Critère	Présent travail	[Kubin et Kleijn, 1999b]	[Feldbauer, 2005]
Bande fréquentielle	[50Hz,7.5kHz]	[20Hz,4kHz]	[100Hz,3.6kHz]
Filtre auditif	cBIT_2^*	GT	GT
Nombre de filtres	32	21	20
τ_s	18%	50%	20%
$\tau_{\text{éch}}$	0.76	1.26	0.9
Qualité de synthèse	Bonne	Bonne	Bonne

cette différence double de valeur et vaut en moyenne -0.5. Pour la valeur de $r = 0.56$, l'ODG moyen pour le groupe des hommes vaut -1.0 alors que celui pour les femmes vaut -1.4. Pour ces valeurs, la qualité de synthèse est plus que moyenne voire bonne (voir tableau 5.1). On attribue cette variation à la différence entre les fréquences des fondamentales

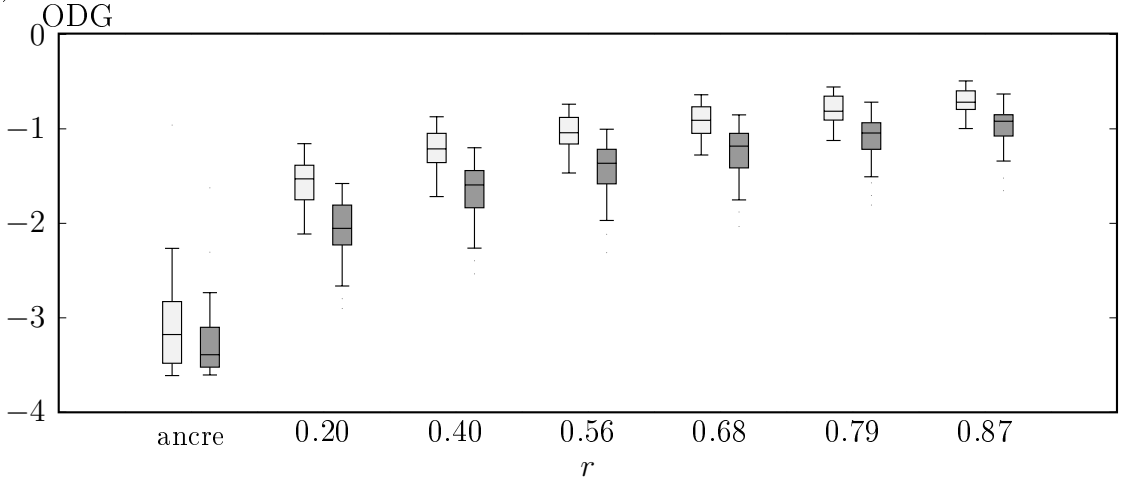


Figure 5.11 ODG moyen de l'ensemble de validation. Les résultats sont séparés par sexe. Les hommes sont représentés par des rectangles clairs alors que les femmes sont représentées par des rectangles foncés.

ainsi qu'à la durée de prononciation des voyelles. Dans [Pépiot, 2015], il a été démontré qu'il existe une différence significative entre les mécanismes intervenant lors la production de la voix entre homme et femme. L'optimisation des paramètres du système par sexe est hors de la portée de cette thèse d'autant plus qu'aucun signal synthétisé score moins que le signal ancre et que la variation des performances du système par sexe est raisonnable.

Les résultats de la figure 5.11 montrent que le modèle du masquage proposé permet de réduire de manière intelligente le nombre d'impulsions à la sortie du modèle neuronal. Le taux de réduction du nombre d'impulsions est reporté par la valeur de τ_s. Pour la valeur de $r = 0.56$, le nombre d'impulsions est réduit par 82% engendrant seulement 0.76 impulsions par échantillon. La dégradation introduite par cette réduction, tel que démontré par les

résultats de la figure 5.11, reste perceptiblement non agaçante. En moyenne la qualité de synthèse est moyenne voire bonne. Ces résultats montrent aussi que l'approche proposée est généralisable et que le facteur r permet de réduire de manière contrôlée le nombre d'impulsions survivant le bloc du masquage.

5.6 Conclusion

Dans ce chapitre on a proposé un nouvel algorithme de masquage dans le domaine perceptuel. Contrairement aux algorithmes publiés dans la littérature, l'approche proposée ne requière pas un stockage de motifs d'excitation précalculés. En effet, on a montré qu'il est possible de calculer les seuils de masquage pré et post-stimuli de façon récursive. Ceci s'est traduit en une complexité d'implémentation réduite. Puisque la mise à zéro des impulsions masquées résulte en une perte d'énergie, un algorithme simple en implémentation de compensation adaptatif a été conçu pour restaurer cette perte. Des signaux de parole ont été utilisés pour valider ces algorithmes. Les résultats présentés dans la section 5.5 montrent qu'il est possible de mettre 82% des impulsions à zéro tout en maintenant une bonne qualité de synthèse. Les motifs d'excitation ainsi obtenus sont épars, et même pour un nombre aussi petit que 0.76 impulsions par échantillon, la transmission de ses motifs requière un débit élevé puisque la position de ces impulsions doit aussi être transmise. Le chapitre suivant présente des algorithmes de compression avec et sans perte dans le but de réduire le débit de transmission de ces motifs.

CHAPITRE 6

Compression des motifs d'excitation auditive

Les motifs d'excitation auditive obtenus par le système proposé sont épars. Quand le modèle de masquage est utilisé pour mettre les valeurs des impulsions masquées à zéro, ces motifs deviennent encore plus épars. Dans le chapitre précédant (section 5.5), on a montré que pour une valeur de $r = 0.56$ mettant 82% des impulsions à zéro il est encore possible de récupérer le signal sans distortions audibles. Habituellement les signaux épars sont représentés par un couple position-amplitude. Dans ce chapitre, on aborde la compression de ces couples dans le but de réduire le débit nécessaire à leur transmission. Alors que la valeur des amplitudes est une variable continue (finement quantifiée avec un quantificateur uniforme de 16 bits), la valeur des positions est une variable entière. Il est important de noter que les erreurs de quantification de ces deux variables sont dépendantes puisque une erreur de quantification de l'amplitude par example est encore plus amplifiée quand il y a une erreur de quantification de sa position. Quand les positions sont compressées sans perte, les valeurs des amplitudes peuvent être compressées avec perte indépendamment de leurs positions. C'est l'approche adoptée dans ce chapitre : les positions des impulsions sont compressées sans perte alors que leurs amplitudes sont quantifiées grossièrement. Dans ce chapitre, on propose des approches permettant de réduire le débit nécessaire à la transmission des motifs d'excitation auditive. On montre dans ce chapitre que la transformation de Burrows-Wheeler conjointement utilisée avec le codage par plage permet de réduire par 80% le débit nécessaire à la transmission des positions des impulsions masquantes. L'encodage des différences entre les amplitudes s'avère être une approche appropriée pour la compression des amplitudes des impulsions puisque cela permet de réduire par 64% le débit nécessaire à la transmission de cette information. Comparativement au débit de 256 kbps [1] nécessaire à la transmission des signaux audio on est capable de réaliser une compression de l'ordre de 65% tout en maintenant une bonne qualité de reconstruction. Ce chapitre ne présente pas un bloc de compression dont la sortie est un flux binaire, mais présente les algorithmes de compression appropriée à la nature des signaux analysés. Pour collecter les mesures subjectives, le bloc de compression compresse et quantifie les motifs d'excitation. Le décodeur opère directement sur ces motifs où le bruit de quantification a été introduit par le codeur.

1. Les signaux audio sont échantillonnés à une fréquence de 16 kHz avec une résolution de 16 bits par échantillon.

6.1 Codage des positions des impulsions

Le bloc du modèle neuronal présenté dans ce travail opère en sélectionnant les maxima après redressement simple-alternance des sorties des filtres du banc d'analyse (voir section 4.1.2). Cette opération génère un train d'impulsions éparse possédant une propriété particulière : Par construction, il n'existe pas deux impulsions successives non séparés par un zéro. Exploitant cette propriété, il est possible d'encoder seulement le nombre de zéros successifs séparant les positions des impulsions masquantes. Dans cette section, on propose des transformations bijectives permettant de réduire l'entropie de l'information décrivant la positions des impulsions. On supporte ces propositions par des expériences supportant la validité de ces approches.

6.1.1 Transformations réversibles du train d'impulsions

Les performances d'un algorithme de compression étant supérieurement bornées par l'entropie de la source, la réduction de cette borne peut se faire en utilisant diverses transformations bijectives. Ceci revient à créer de la redondance exploitable [Lakhdhar, 2009; Lakhdhar et Lefebvre, 2012]. Une méthode des plus simples consiste à encoder le nombre de répétitions quand l'alphabet est réduit et la séquence à transmettre est relativement longue (codage par plage (*run length coding*) (RLE)). Pour que l'algorithme RLE soit efficace, il est préférable que les symboles identiques soient successifs. La transformation de Burrows-Wheeler (BWT) [Burrows et J. Wheeler, 1994] est une transformation réversible qui permet d'augmenter la probabilité que des symboles initialement éloignées les uns des autres se trouvent côte à côte. Considérons l'example de la séquence suivante : $S = [1, 0, 1, 0]$. La transformation consiste à trier par ordre lexicographique les rotations de la séquence S, le résultat final étant la dernière colonne $S' = [1, 1, 0, 0]$ dont le codage par plage est plus efficace que ce celui de S.

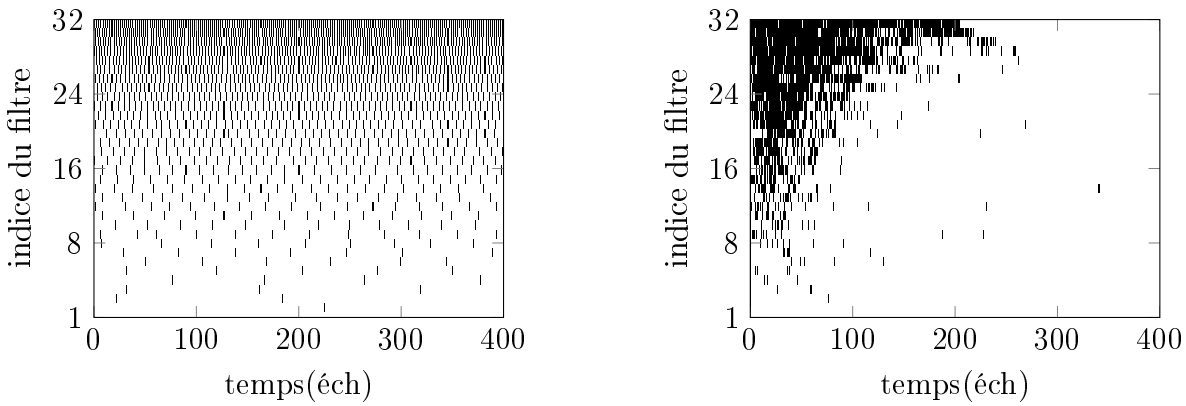
									↓		
1.	1	0	1	0		2.	0	1	0	1	
2.	0	1	0	1		4.	0	1	0	1	
3.	1	0	1	0		1.	1	0	1	0	←
4.	0	1	0	1		3.	1	0	1	0	

Pour récupérer la séquence originale, le décodeur a besoin évidement de la séquence transformée mais aussi de l'indice i de S dans la table des rotations (illustré dans l'example donné par la flèche horizontale). Le décodeur opère de manière itérative où à chaque étape la séquence obtenue est triée par ordre lexicographique ensuite la séquence S' y est jux-

taposée. Cette étape est répétée jusqu'à ce que la longueur des symboles juxtaposés soit égale à la longueur de la séquence S'. La séquence S est tout simplement récupérée en utilisant l'indice i.

1	10	101	1010	0101	
1	10	101	1010	0101	
0	01	010	0101	1010	←
0	01	010	0101	1010	

La transformée de Burrows-Wheeler est une transformation bijective qui permet d'augmenter la probabilité que les symboles identiques se trouvent groupés. Cette propriété s'est avérée très efficace pour accélérer l'opération d'indexage et séquençage de l'ADN [Li et Durbin, 2009].



(a) Positions des impulsions non-nulles d'une trame d'un signal de parole.

(b) Positions des impulsions non-nulles après application de la transformation de Burrows-Wheeler.

Figure 6.1 Example d'application de la transformation de Burrows-Wheeler. Les positions des impulsions non-nulles sont marquées par la couleur foncée.

La figure 6.1 donne un example de l'application de cette transformation **bijective** sur les positions des impulsions non nulles d'une trame d'un signal de parole. Alors que ces positions sont dispersées sur la figure 6.1(a), l'application de la transformation BWT permet de regrouper les positions non nulles ensemble tel qu'illustré sur la figure 6.1(b). Il est évident que cette transformation permet d'améliorer les performances de l'algorithme RLE puisqu'elle a tendance à grouper les symboles identiques ensemble [2].

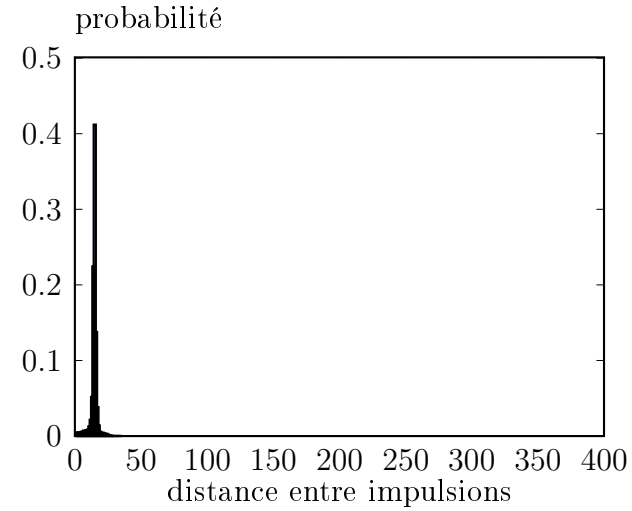
2. Les symboles ici sont tout simplement 1 et 0 signifiant respectivement présence et absence d'impulsion.

La figure 6.2 donne la distribution des distances séparant les impulsions non nulles du filtre auditif centré autour de 1kHz. Sur la figure 6.2(a), cette distribution est mono-modale puisque la distance entre impulsions non nulles est souvent égale à l'inverse de la fréquence centrale du filtre auditif concerné. Quand le nombre des impulsions est réduit grâce à l'application du seuil de masquage, la distribution devient multi-modale (figure 6.2(c)). L'application de la transformation de BWT dans les deux cas permet de transformer ces distributions en distributions où la probabilité des distances entre impulsions décroît exponentiellement. Cette transformation permet aussi de réduire le nombre de distances à transmettre par trames : sur la figure 6.2(b) et la figure 6.2(d) on note la présence de distance valant quasiment la longueur de la trame d'analyze (400 échantillons) ce qui suggère une réduction du nombre de symbole à transmettre d'où la réduction du débit moyen.

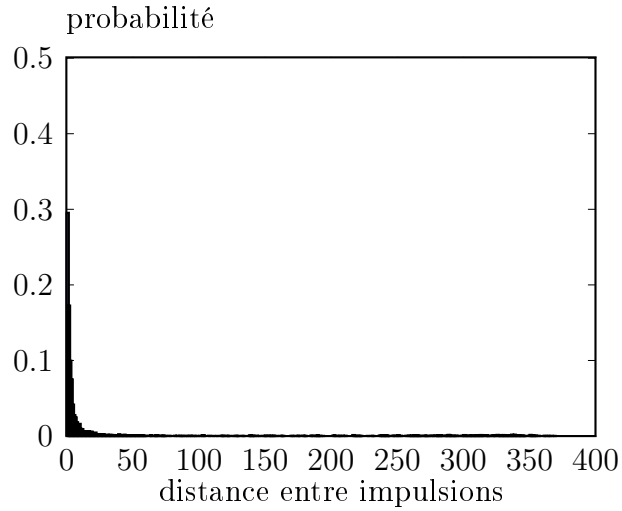
6.1.2 Résultats expérimentaux

La figure 6.3 donne le nombre de bits/symbole ainsi que le débit total nécessaire à la transmission de la position des impulsions non nulles avant et après application du seuil de masquage. Dans chaque sous-figure trois courbes sont données. La première représentée par le symbole triangle, donne ces quantités quand aucune transformation n'est appliquée : les positions des impulsions sont transmises telles quelles sous forme de successions de 1 et 0. La deuxième courbe représentée par le symbole + décrit les mêmes quantités quand l'algorithme RLE est appliqué au train d'impulsions et seule la distance entre positions d'impulsions non nulles est transmise. La courbe représentée par le symbole carré donne le nombre de bits/symbole ainsi que débit total quand la transformation BWT est appliquée avant l'utilisation de l'algorithme RLE. Les données sont collectées à partir des motifs d'excitation générés suite au passage des signaux audio donnés en annexe A à travers le système proposé.

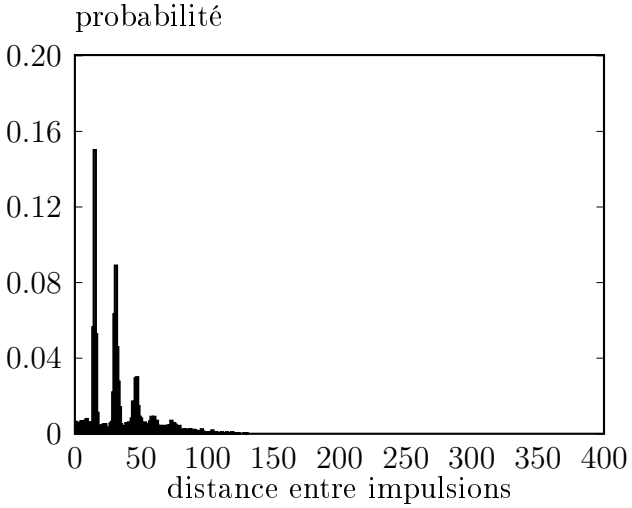
La figure 6.3(a) montre que l'entropie du train d'impulsions complet croît avec la fréquence centrale du filtre auditif. Ceci s'explique par le fait que la distance entre impulsions est proportionnelle à l'inverse de la fréquence du filtre auditif (voir section 5.3 équation 4.7). Pour les fréquences voisinant par example la moitié de la fréquence d'échantillonnage la probabilité d'avoir ou non une impulsion vaut 0.5 et par conséquent il est nécessaire d'encoder cette information en utilisant 1 bit/position. Quand la transformation RLE est appliquée à ce train d'impulsions l'entropie par symbole (i.e. distance séparant deux impulsions consécutives) vaut en moyenne 3.2 bits/distance. L'application de la transformation BWT augmente cette quantité surtout vers les hautes fréquences. En effet vers les



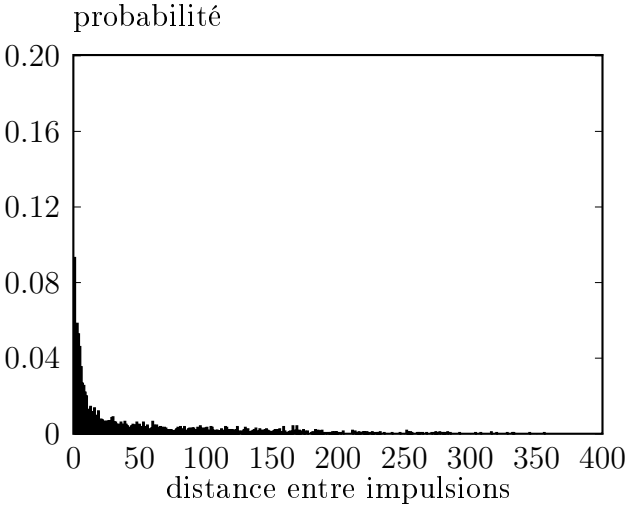
(a) Probabilité des distances entre impulsions avant masquage.



(b) Probabilité des distances entre impulsions avant masquage et après application de la transformation de BWT.

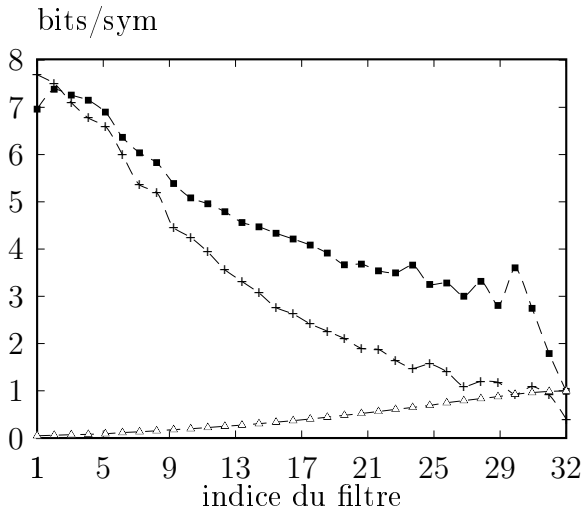


(c) Probabilité des distances entre impulsions après masquage.

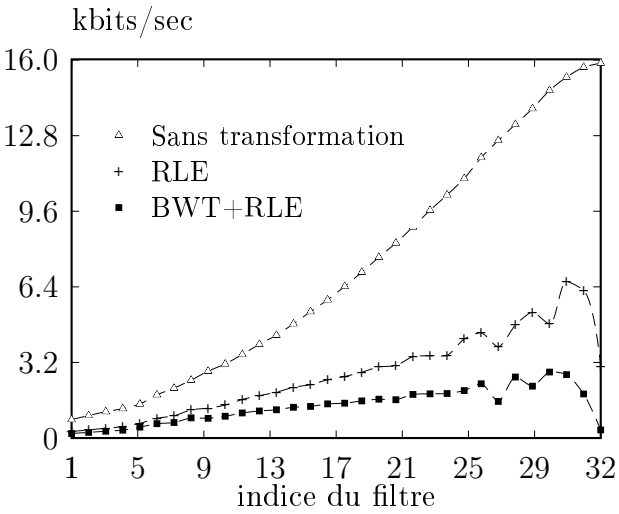


(d) Probabilité des distances entre impulsions après masquage et application de la transformation de BWT.

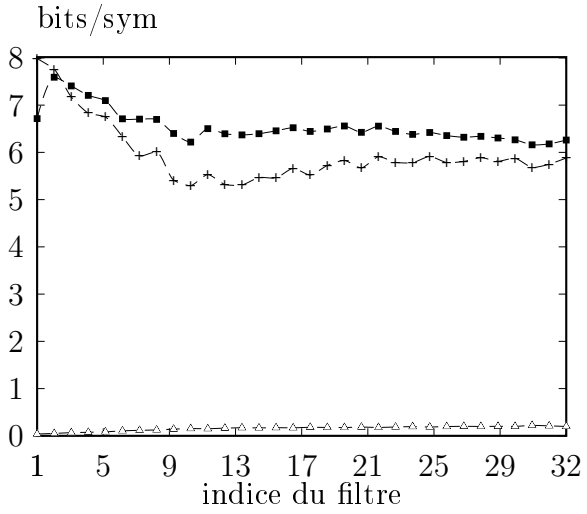
Figure 6.2 Probabilité des distances entre impulsions avant et après application du masquage avec ou sans la transformation de BWT pour le filtre auditif centré autour de 1kHz.



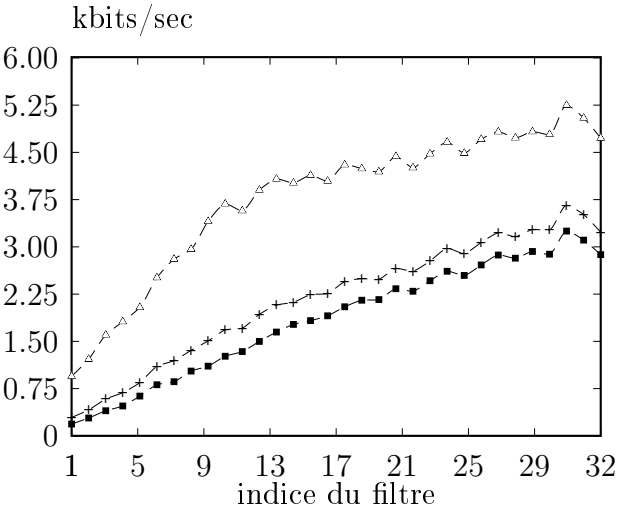
(a) Nombre de bits/symbole nécessaire pour la transmission de la position des impulsions.



(b) Débit binaire nécessaire pour la transmission de la position des impulsions.



(c) Nombre de bits/symbole nécessaire pour la transmission de la position des impulsions après application du masquage.



(d) Débit binaire nécessaire pour la transmission de la position des impulsions après application du masquage.

Figure 6.3 Nombre de bits/symbole et débit total nécessaire à la transmission des positions des impulsions avec et sans compression avant et après application du seuil de masquage. Le triangle représente ces quantités quand aucune transformation n'est utilisée. Le signe + représente cette information quand le RLE est utilisé et le carré la représente quand le RLE et utilisé conjointement avec la transformation BWT.

hautes fréquences, puisque quasiment la distance entre impulsions vaut 1 échantillon, la probabilité d'avoir une succession de 1 suivie d'une succession de 0 de mêmes longueurs augmente (voir figure 6.1). Même si les entropies par symbole augmentent suite à ces transformations, la longueur de la séquence à transmettre diminue ce qui fait que le débit moyen pour la transmission des positions des impulsions diminue. La figure 6.3(b) donne le débit moyen pour la transmission de l'information décrivant la positions des impulsions. Même si l'entropie des impulsions sans transformations est inférieure à 0.5 bit/symbole en moyenne, transmettre cette information requière un débit élevé puisque la séquence à transmettre est longue et cette opération consiste à transmettre même les positions des impulsions nulles. Le débit moyen quand seulement la distance entre impulsions non nulles est transmise est plus petit. En effet, exploiter la parcimonie du train binaire s'avère justifiée puisque le débit moyen total est réduit par 62%. Quand la transformation BWT est appliquée avant l'utilisation de l'algorithme RLE pour encoder la distance entre impulsions, le débit moyen total est réduit par 82% (voir tableau 6.1). L'utilisation de la transformation BWT conjointement à l'algorithme RLE permet de réduire le débit par moitié comparativement à l'utilisation de l'algorithme RLE.

Quand le seuil de masquage ($r = 0.56$) est appliqué au même train d'impulsions, les constations sont différentes. L'entropie donnée en bits/impulsion sur la figure 6.3(c) nécessaire à la transmission des positions sans transformation se trouve réduite puisque dans ce cas la probabilité d'avoir une impulsion non nulle devient petite (l'application du seuil de masquage avec la valeur de r choisie résulte en une mise à zéro de 82% des impulsions comme décrit dans la section 5.5.1). On note aussi que l'entropie des distances séparant les impulsions masquantes avec ou sans la transformation BWT est quasiment la même et vaut en moyenne 6 bits/distance. Le bénéfice généré par l'utilisation de la transformation BWT se traduit par une séquence plus courte à transmettre. La figure 6.3(c) illustre cela : la longueur moyenne et par la même occasion le débit moyen sont réduits par 15% comparativement au cas où seul l'algorithme RLE est utilisé pour transformer les positions en distances. Le tableau 6.1 résume les résultats donnés sur la figure 6.3. L'utilisation

Tableau 6.1 Débit nécessaire à la transmission des positions des impulsions.

Transformation	Train d'impulsions complet	Train d'impulsions réduit ($r = 0.56$)
Sans transformation	220.90 kbit/s	78.57 kbit/s
RLE	83.27 kbit/s	45.40 kbit/s
BWT+RLE	42.10 kbit/s	38.47 kbit/s

conjointe du masquage combiné à la transformation BWT suivie par l'encodage RLE réalise le meilleur taux de compression. En effet, on est capable de réduire par un facteur de

82% le débit total pour la transmission des positions des impulsions. Puisque c'est une compression sans pertes, la qualité de reconstruction se trouve inchangée.

6.1.3 Discussions

On a présenté dans cette section les approches possibles permettant de réduire le nombre de bits nécessaires à la transmission des positions des impulsions masquantes. On a montré qu'il est possible d'exploiter la parcimonie de cette information en applicant des transformations réversibles réduisant encore son entropie. Étant donné l'alphabet réduit (1 pour indiquer la présence d'une impulsion, 0 pour indiquer son absence) la transformation BWT s'est avérée très efficace puisqu'elle permet de réduire par moitié le débit moyen nécessaire pour la transmission des postions des impulsions. Quand le seuil de masquage est appliqué, le bénéfice généré par cette transformation est moins important mais on est capable de réduire par un facteur de 15% le débit moyen comparativement à l'application de l'algorithme RLE. D'autres transformations/algorithmes de compression sans perte ont été évalués mais exclus soit pour moindres performances ou pour des raisons de complexité d'implementation. Le codage par dictionnaire adaptatif n'est efficace que quand une source génère des séquences formées par des blocs dont le nombre est réduit [Ziv et Lempel, 1978]. L'analyse de l'autocorrelation entre les distances séparant les impulsions montre que cette approche n'est pas efficace pour l'encodage des distances séparant les impulsions masquantes. La figure 6.4 illustre l'autocorrélation des distances séparant les

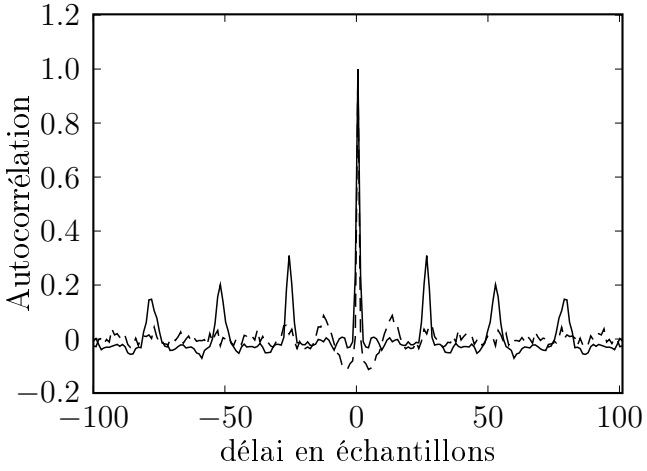


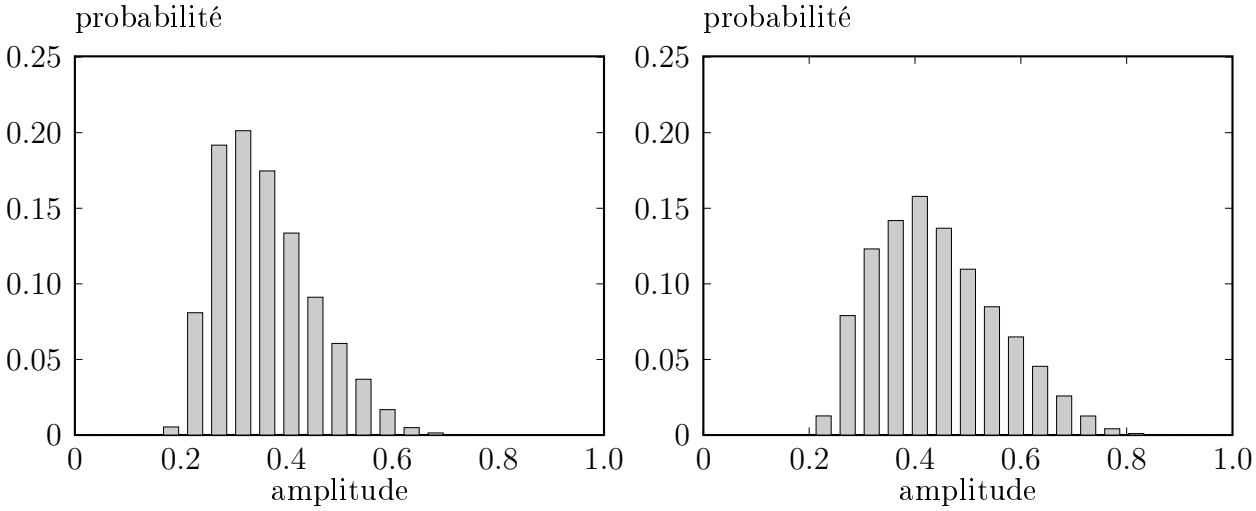
Figure 6.4 Autocorrélation entre distances séparant les impulsions masquantes pour le filtre auditif centré autour de 1kHz avant (ligne continue) et après (ligne discontinue) application de la transformation BWT.

impulsions masquantes. Après application de la transformation BWT, les distances deviennent non corrélées et donc une compression par dictionnaire ou par prédiction n'est

plus justifiée. Cette constatation valide le choix de la transformation BWT comme une transformation simple, élégante et efficace pour réduire le débit moyen nécessaire pour la transmission des positions des impulsions masquantes.

6.2 Codage des amplitudes des impulsions

La figure 6.5 donne la probabilité des amplitudes des impulsions avant et après application du masquage. L'application du masquage transforme la distribution des amplitudes des impulsions. En comparant les figures 6.5(a) et 6.5(b) on constate que les valeurs extrêmes des amplitudes sont plus élevées après masquage. La valeur moyenne des amplitudes des impulsions elle aussi augmente. L'algorithme de correction adaptative en effet amplifie les amplitudes des impulsions masquantes pour compenser la perte d'énergie résultante de la mise à zéro des impulsions masquées (voir section 5.3) . Cette correction a pour effet de transformer la distribution des amplitudes des impulsions qui devient plus proche d'une distribution uniforme impliquant une entropie plus élevée. Ceci est illustré sur figure 6.6

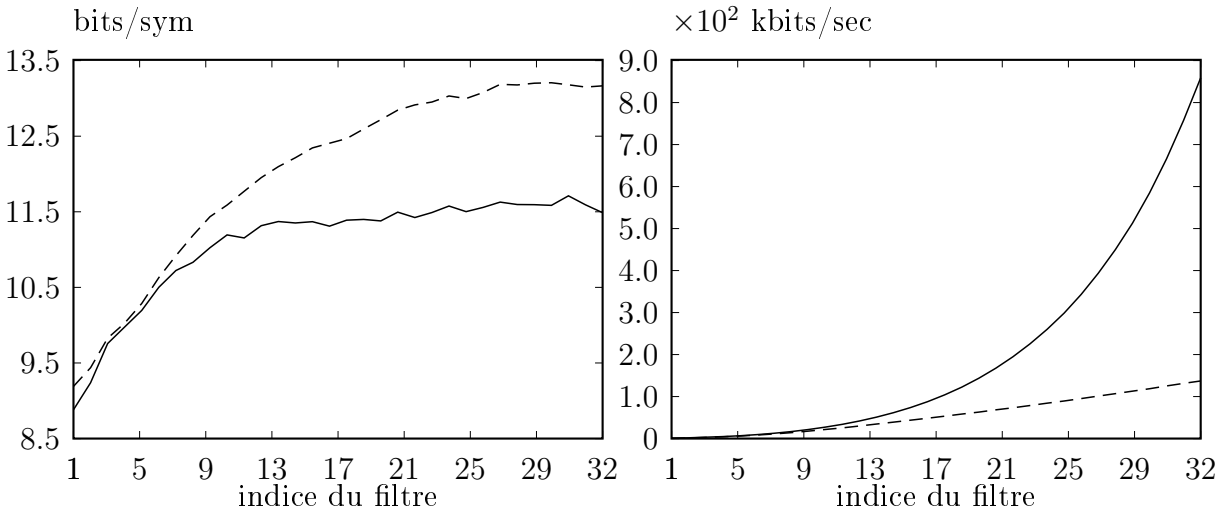


(a) Probabilité des amplitudes des impulsions avant masquage.

(b) Probabilité des amplitudes des impulsions après masquage.

Figure 6.5 Probabilité des amplitudes des impulsions avant et après application du masquage pour $r = 0.56$.

où l'entropie moyenne des amplitudes des impulsions augmente après application du masquage. Même si l'entropie des amplitudes augmente dû à l'application du masquage, la longueur moyenne du code nécessaire à leurs transmissions diminue puisque pour $r = 0.56$, 82% des impulsions sont mises à zéro. La figure 6.6(b) illustre le débit moyen cumulatif nécessaire à la transmission de cette information. Si les amplitudes sont transmises sans



(a) Entropie moyenne des amplitudes des impulsions masquantes.

(b) Débit moyen cumulatif nécessaire pour la transmission des amplitudes des impulsions.

Figure 6.6 Entropie et débit moyens des amplitudes des impulsions avant (ligne continue) et après (ligne discontinue) application du masquage pour $r = 0.56$.

pertes il faut en moyenne 820kbits/sec avant masquage contre 133 kbits/sec après application du masquage. Il est à noter que contrairement au cas sans masquage, le débit augmente linéairement avec le nombre des filtres auditifs (en moyenne 5 kbits/sec/filtre). L'application du masquage permet de réduire par 84% le débit nécessaire à la transmission sans perte des amplitudes des impulsions. Cependant, des expérimentations conduites sur les signaux de parole utilisés dans ce travail ont révélé que la quantification des valeurs des amplitudes des impulsions telles quelles ne donne pas des résultats satisfaisants quand le débit de transmission est réduit. On propose de modéliser ces valeurs par un modèle moyenne mobile, *moving average* (MA) dans le but d'exploiter la redondance présente dans ces amplitudes.

6.2.1 Modélisation des amplitudes

La figure 6.8 présente l'autocorrélation entre amplitudes des impulsions quand on fait abstraction de leurs positions : on concatène les valeurs des impulsions non nulles en respectant l'ordre dans lequel elles sont générées. La figure 6.8(a) présente l'autocorrélation entre valeurs d'amplitudes d'impulsions masquantes consécutives. Il est clair que les amplitudes des impulsions consécutives sont fortement corrélées. On envisage alors la possibilité de transmettre l'erreur de prédiction à la place de transmettre la valeur de l'amplitude elle-même. Cette technique est souvent utilisée pour le codage des signaux de parole [Bes-

sette *et coll.*, 2002; Jayant, 1974; Ning et Deriche, 2003; Salami *et coll.*, 1998; Schroeder et Atal, 1985].

Sur la figure 6.8(b) l'erreur moyenne de prédiction est donnée. Cette erreur est calculée comme étant l'erreur de prédiction quand un modèle à MA est utilisé pour prédire la valeur des amplitudes des impulsions à partir de celles précédentes. Si v_n représente la valeur de l'amplitude masquante à la position n, alors l'estimée de la valeur v_{n+1} est donnée par :

$$v_{n+1} = v_n + \sum_{i=1}^{N} \alpha_i v_{n-i} + \epsilon_n \tag{6.1}$$

Si on suppose que $N = 1$ et on fixe $\alpha_i = 1$, il est possible de déduire une condition suffisante, quand vérifiée, présente un avantage considérable à la simple quantification de v. L'erreur quadratique moyenne de prédiction $\mathbf{E}[\epsilon_n^2]$ peut être estimée :

$$\begin{aligned}\mathbf{E}[\epsilon_n^2] &= \mathbf{E}[(v_{n+1} - v_n)^2] \\ &= \mathbf{E}[v_{n+1}^2] + \mathbf{E}[v_n^2] - 2\mathbf{E}[v_{n+1}v_n] \\ &= \mathbf{E}[v_n^2][2(1 - C_1)]\end{aligned}$$

Où C_1 représente la corrélation normalisée entre valeurs d'amplitudes adjacentes. Si la valeur de C_1 est supérieure à 0.5, alors la valeur de $\mathbf{E}[\epsilon_n^2]$ est plus petite de celle de $\mathbf{E}[v_n^2]$. Puisque l'erreur de quantification dépend de la variance du signal à l'entrée du quantificateur, l'utilisation de ϵ à l'entrée du quantificateur permet la réduction du nombre de bits nécessaire à la transmission des valeurs des amplitudes des impulsions. De façon générale, pour $N = 1$:

$$\mathbf{E}[\epsilon_n^2] = \mathbf{E}[v_n^2][1 + \alpha_1^2 - 2\alpha_1 C_1] \tag{6.2}$$

Cette quantité est minimale quand $\alpha_1 = C_1$. Il est important de noter que l'implémentation d'une telle approche requière une forme de rétro-action de telle sorte qu'il n'y a pas d'accumulation d'erreurs au niveau du décodeur. Ceci peut être réalisé par la quantification de la valeur de $\epsilon_n + \epsilon_{n-1} - \hat{\epsilon}_{n-1}$ où $\hat{\epsilon}_{n-1}$ est la valeur quantifiée de la différence à l'instant $n - 1$ [3]. Cette approche peut être implémentée en utilisant un codeur opérant par codage par modulation des différences, *Differential pulse code modulation* (DPCM). La figure 6.7 donne un schéma d'une implémentation possible d'une telle approche. La loupe de rétro-action assure que l'erreur entachant le signal reconstruit v est celle due à la quantification de ϵ_n et non une accumulation d'erreurs dues à la quantification des valeurs

3. La valeur quantifiée d'une variable x est notée $\hat{x}$.

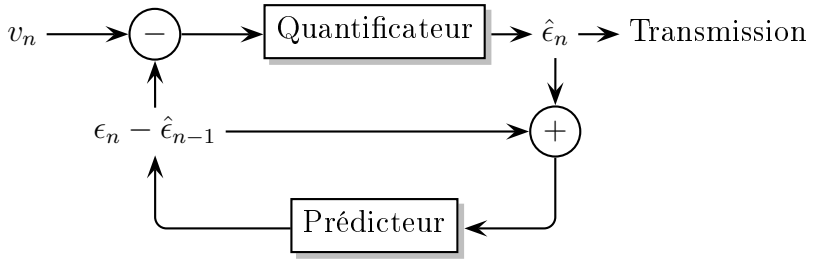


Figure 6.7 Codage des différences entre valeurs des amplitudes masquantes.

v_n précédentes.

$$\epsilon_n - \hat{\epsilon}_n = v_n - \alpha_1 \hat{v}_n - \hat{\epsilon}_n = v_n - \hat{v}_n \tag{6.3}$$

6.2.2 Résultats expérimentaux

Dans la figure 6.8(b), pour chaque valeur de l'ordre N du modèle MA l'erreur ϵ décrite dans l'équation (6.1) est donnée en dB. Les symboles clairs représentent le cas où un modèle différent est utilisé par filtre auditif. Le cas où le même modèle est utilisé pour tous les filtres auditifs est représenté par la couleur foncée. Avant l'application du seuil de masquage, l'erreur de prédiction (représenté par le symbole carré sur la figure 6.8(b)) vaut en moyenne -11 dB. L'augmentation de l'ordre du modèle MA ne permet pas de la diminuer. Le bénéfice engendré par l'utilisation de modèles MA par filtre est minimal. Quand l'ordre du modèle MA est fixé à 1, ce gain est nul. On constate le même phénomène après application du seuil de masquage : l'erreur de prédiction est sensiblement la même peu importe si on utilise un seul modèle MA ou plusieurs. Cependant, quand l'ordre du modèle passe de 1 à 2, l'erreur de prédiction décroît par 1dB. Pour un ordre encore plus élevé aucune amélioration ne peut être observée. Pour ces raisons, l'ordre du modèle MA utilisé est fixé à 1. Vu qu'il n'y pas de gain à modéliser indépendamment les valeurs des amplitudes des impulsions par filtre, le modèle suggéré est identique pour tous les filtres auditifs.

Dans le cas où $r = 0.56$, on peut écrire que :

$$v_{n+1} = v_n - 0.9933 \times v_{n-i} + \epsilon_n \tag{6.4}$$

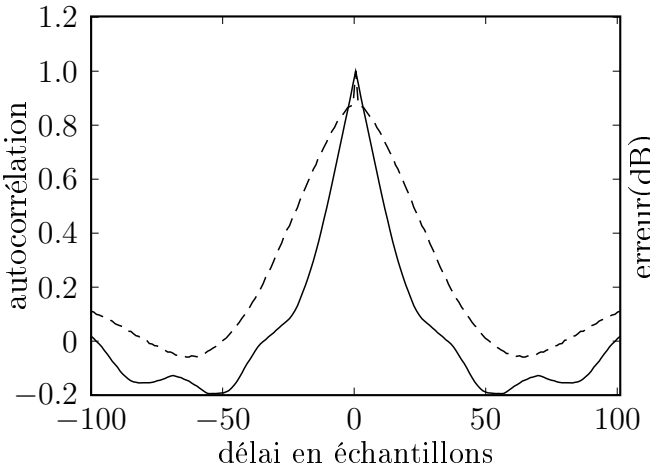
Le modèle donné par l'équation (6.4) est statique et ne requière donc pas de la transmission d'informations secondaires au décodeur.

Avec cette valeur, les distributions des erreurs des prédictions ϵ_n sont données sur la figure 6.8(c) et la figure 6.8(d). À comparer avec la figure 6.6, la dynamique de l'information à transmettre est réduite grandement. On remarque aussi que la valeur de $\epsilon_n = 0$ est

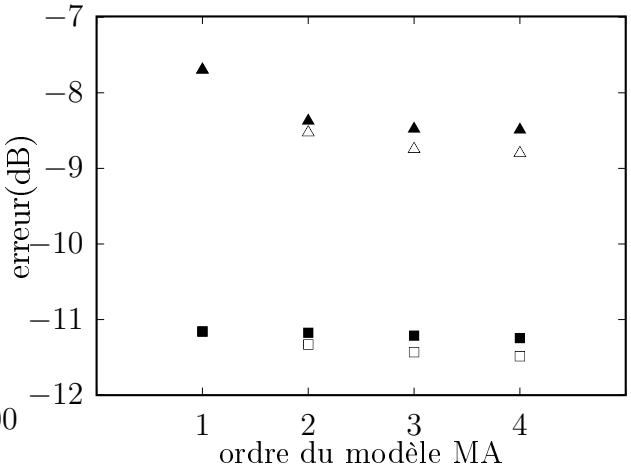
fortement présentée : la moitié des valeurs des amplitudes des impulsions masquantes peut être prédite avec une erreur quasiment nulle par un modèle MA d'ordre 1. Quand le seuil de masquage est appliqué, l'erreur de prédiction est moins centrée autour de zéro mais possède quand même une dynamique moins élevée que celle présentée sur la figure 6.5(b).

La figure 6.9(a) représente l'erreur de quantification des valeurs des amplitudes des impulsions masquantes pour les deux approches investiguées dans cette section. La première approche consiste à quantifier ces valeurs directement alors que la deuxième consiste à utiliser le codeur illustré sur la figure 6.7 où la différence entre valeurs d'amplitudes d'impulsions successives est quantifiée. Les signaux audio de l'annexe A sont utilisés pour produire ces résultats. Les motifs d'excitation auditives sont extraits à partir des ces signaux. Ces motifs sont ensuite subdivisés en deux ensembles : le premier comptant 10% des données est utilisé pour concevoir et entraîner des quantificateurs scalaires, le deuxième est utilisé pour estimer leurs performances. Pour un nombre de bits donné, l'algorithme de Lloyd [Lloyd, 1982] est utilisé pour choisir les pas de quantification optimaux utilisant les données du premier ensemble. Ce quantificateur, une fois conçu, est utilisé par la suite pour quantifier les données du deuxième ensemble et estimer l'erreur objective de quantification.

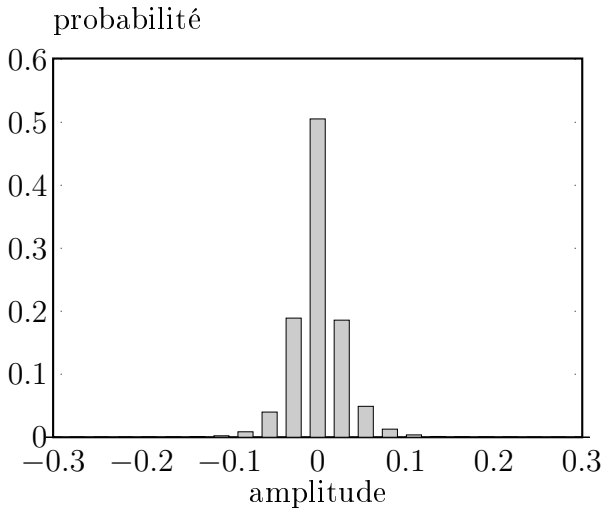
La quantification des différences nécessite en moyenne 1 bit en moins pour la même qualité objective de reconstruction. Cette approche est donc préférée à celle où les valeurs des amplitudes des impulsions sont quantifiées telles quelles. La figure 6.9(b) donne l'erreur de quantification en fonction du débit exprimé en bit/sec. Ce débit est calculé comme étant le nombre de bits par quantificateur multipliés par le nombre d'impulsions par filtre auditif et est exprimé en secondes. Pour un débit donné par quantificateur, l'erreur objective de reconstruction suit la même forme : elle décroît linéairement en fonction du débit total (fréquence centrale du filtre auditif) ensuite sature. Ceci prouve que l'allocation du même débit par filtre auditif ne permet pas d'avoir une erreur de reconstruction objective distribuée uniformément à travers les fréquences. Dans le but de distribuer le bruit de quantification uniformément à travers la bande de fréquence couverte par le banc de filtres, 1 bit additionnel a été alloué aux filtres auditifs dont la fréquence centrale est inférieure à 2kHz. La taille du quantificateur a été variée d'un pas de 1 bit et pour chaque configuration les filtres auditifs des basses fréquences ont jouit d'un bit additionnel. Par example, quand 1 bit est utilisé pour les quantificateurs des filtres dont les fréquences sont supérieures à 2kHz, les filtres restant utilisent des quantificateurs à deux bits chacun. Les signaux audio de l'ensemble d'entraînement ont été encodés et leurs motifs d'excitation ont étés utilisés pour



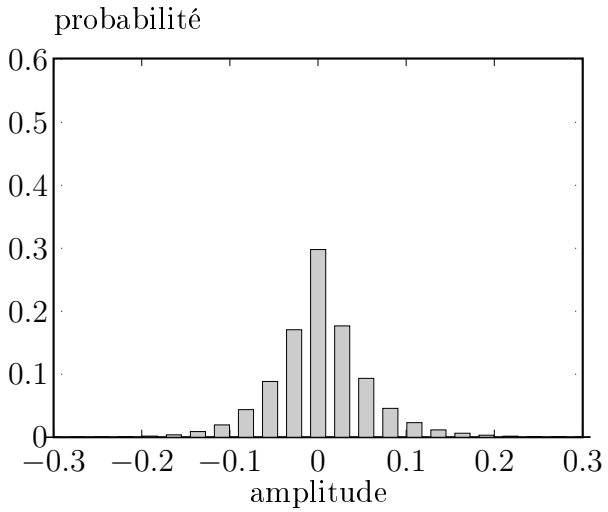
(a) Autocorrélation des amplitudes des impulsions masquantes des filtres auditifs centrés autour de 250Hz (ligne discontinue) et 4.2 kHz (ligne continue).



(b) Erreur de modélisation des amplitudes des impulsions masquantes pour différents ordres du modèle linéaire.



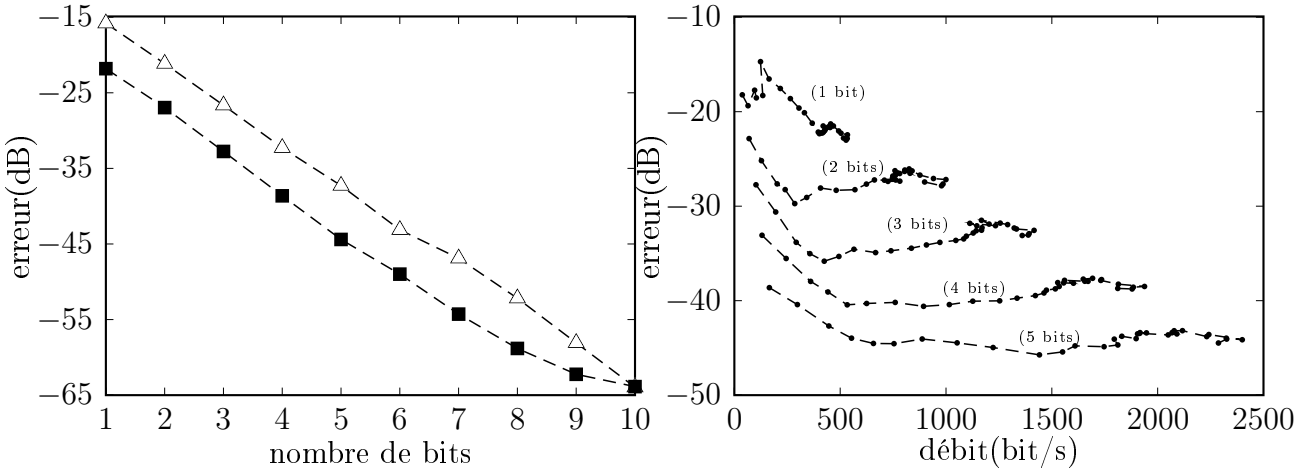
(c) Probabilité d'erreur de prédiction des valeurs des amplitudes des impulsions avant masquage.



(d) Probabilité d'erreur de prédiction des valeurs des amplitudes des impulsions après masquage.

Figure 6.8 Autocorrélations des amplitudes des impulsions masquantes et erreur de prédiction. Sur la figure 6.8(b), les erreurs de prédiction des amplitudes des impulsions avant et après masquage ($r = 0.56$) sont respectivement représentées par le symbole carré et triangle. Les couleurs claires et foncées représentent cette erreur quand un modèle MA est utilisé par filtre auditif et quand le même modèle est utilisé pour tous les filtres auditifs respectivement.

concevoir et optimiser les différents quantificateurs. Les motifs d'excitation auditive des signaux de validation ont étés quantifiées utilisant ces quantificateurs et utilisés par la suite pour estimer la valeur de l'ODG moyen. La figure 6.10 donne l'ODG moyen de l'ensemble



(a) Erreur de prédiction pour différents nombre de bits. Les erreurs de prédiction avant et après masquage ($r = 0.56$) sont respectivement représentées par le symbole carré et triangle.

(b) Erreur de quantification en fonction du débit donnée par filtre auditif. La taille du quantificateur scalaire est donnée entre parenthèses.

Figure 6.9 Erreur de prédiction pour différents nombre de bits. L'approche basée sur la quantification des différences est représentée par le symbole carré. Les points sur la figure 6.9(b) représentent différent filtres auditifs.

des signaux de validation en fonction du débit utilisé pour la transmission des valeurs des impulsions masquantes quand $r = 0.56$. Vers les bas débits, l'ODG croît linéairement

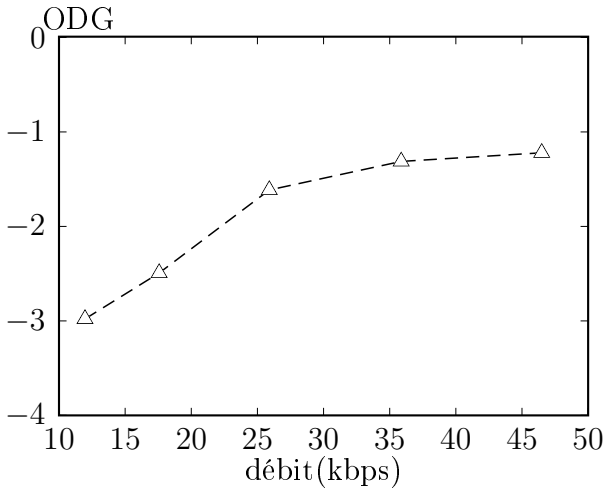


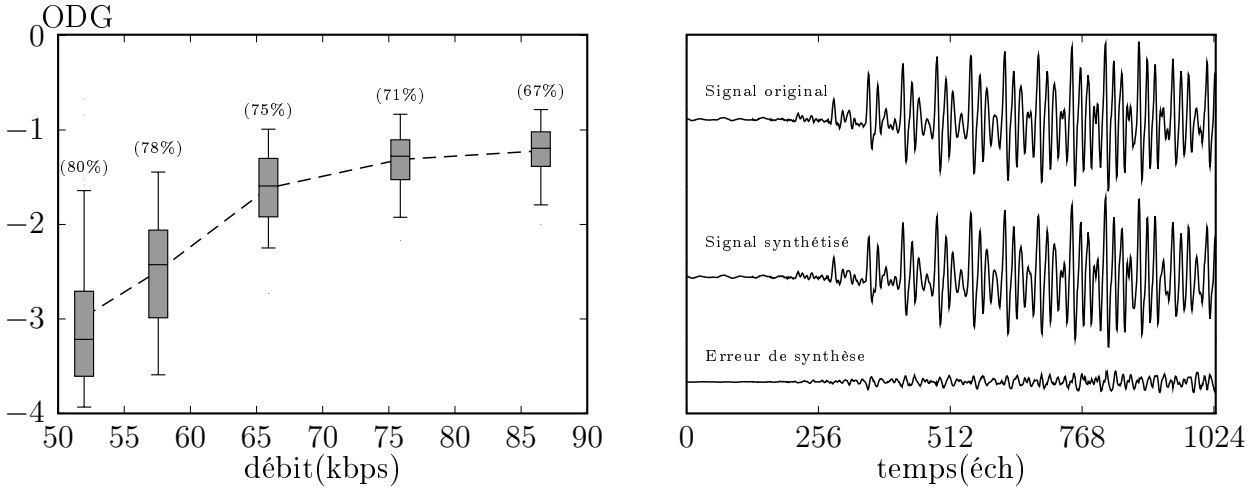
Figure 6.10 ODG moyen en fonction du débit de quantification des valeurs des impulsions masquantes.

puis sature à partir de 30 kbps. À 47 kbps la qualité subjective de synthèse est égale à celle obtenue sans compression des amplitudes et vaut -1.25. Comparativement au débit total sans compression cela revient à une compression de l'ordre de 64% sans dégradations

subjectives. Si on tolère une légère dégradation [4], à 26 kbps le taux de compression vaut 80%.

6.2.3 Discussions

Dans ce chapitre on a présenté les approches permettant de compresser les motifs d'excitation auditive. On a montré dans la section 6.1 que la transformation de BWT utilisée conjointement à la transformation RLE permet de réduire considérablement le débit nécessaire à la transmission des positions des impulsions. Dans la section 6.2 on a montré que la compression avec pertes des différences entre amplitudes est une approche viable permettant de réduire le débit nécessaire à la transmission de cette information. L'application conjointe de ces deux approches permet de réduire le débit total nécessaire à la transmission des motifs d'excitation auditive. La figure 6.11 donne l'ODG moyen en fonction du débit total nécessaire à la transmission des motifs d'excitation auditive. Un example de synthèse d'un signal audio à partir de sa représentation compressée est donné sur la figure 6.11(b). Pour un débit moyen de 76 kbps la qualité de synthèse est bonne.



(a) ODG moyen en fonction du débit de transmission des motifs d'excitation auditive. Le taux de compression est donné entre parenthèse.

(b) Example d'une trame de signal de parole encodée à 76 kbps.

Figure 6.11 ODG moyen en fonction du débit de compression des motifs d'excitation auditive.

Avec ces valeurs on peut alors compléter le tableau 5.2. On a aussi inclus les résultats des travaux de [Thiemann, 2011] où son approche est basée sur le codeur de [Feldbauer, 2005]. La complexité d'implémentation est aussi donnée sur le même tableau.

4. Avec une valeur d'ODG de -1.65 la qualité de reconstruction reste bonne (voir tableau 5.1).

Tableau 6.2 Comparaison entre différents systèmes de synthèse de signaux de parole à partir de leurs motifs d'excitation auditive.

	Bande étroite		Large bande	
Critère	[Kubin et Kleijn, 1999b]	[Feldbauer, 2005]	[Thiemann, 2011]	Présent travail
Bande fréq.	[20Hz,4kHz]	[100Hz,3.6kHz]	[100Hz,7kHz]	[50Hz,7.5kHz]
Filtre auditif	GTF	GTF	GTF	BIT
N. de filtres	21	20	65	32
Parcimonie	50%	20%	20%	18%
Imp./Éch	1.26	0.91	0.53	0.76
Débit (kbps)	70	NA	50	76
Complexité	Réduite	Élevée	Élevée	Réduite
Qualité	Bonne	Bonne	Mauvaise	Bonne

Comme discuté dans la section 5.5, cette comparaison est approximative puisque ces résultats dépendent fortement de la nature des signaux analysés (voir section 5.3.4 de [Thiemann, 2011]). Dans ce travail, cependant, les résultats présentés sont statistiquement signifiants (section 5.5) puisqu'une variété de signaux est utilisée. On s'est aussi assuré qu'il n'y a pas de recouvrement entre l'ensemble d'entraînement et celui de validation. Il faut aussi noter la différence entre les bandes fréquentielles couvertes par les différents codeurs ainsi que leurs nombre de filtres. Ceci complique la comparaison, mais de façon générale le tableau 6.2 montre que l'approche proposée se distingue par une meilleure qualité de reconstruction avec une moindre complexité de reconstruction. Dans [Feldbauer, 2005] des approches basées sur la quantifications vectorielles conjointes des positions et amplitudes avec et sans perte ont étés proposées. Cependant, l'absence de critères corrélant la distortion objective à celle subjective n'est pas évidente. On avait essayé la quantification vectorielle des valeurs des amplitudes des impulsions masquantes et on s'est heurté à plusieurs difficultés. La plus évidente est celle concernant la formation des vecteurs à partir des trains d'impulsions épars. Outre la taille variable de ces vecteurs par trames (dépendemment du signal analysé), la synchronisation de ces vecteurs pour que leurs valeurs maximales résident dans la même dimension requière la modification de leurs positions. Nous avons quand même essayé d'implémenter ces approches mais les résultats obtenus sont moins bons que ceux reportés dans cette thèse. L'approche par codage des différences contourne ces problèmes et peut être considérée comme un quantificateur vectoriel à deux dimensions.

6.2.4 Complexité computationnelle de l'implémentation

L'implémentation des algorithmes proposés dans cette thèse visait en premier lieu l'étude de la faisabilité d'une approche par inversion des motifs d'excitation auditive. De ce fait, le code a été écrit en language de programmation haut niveau [MatLab®, 2011] facilitant

ainsi l'implémentation et l'expérimentation. Ceci a conduit à une redondance dans les traitements ainsi qu'au stockage de données intermédiaires nécessaires à la création des figures par example mais inutiles au fonctionnement du système proposé. La version en C de l'algorithme de détection des pics cependant a été utilisée à la place de son homologue en MatLab (*findpeaks*) dont l'execution est très lente [Yeara Kozlov, 2013][5].

Le tableau 6.3 donne la vitesse d'exécution des principaux algorithmes proposés exprimée comme étant le rapport entre la durée de leurs executions et la durée du signal analysé. Souvent ce rapport est reporté comme un facteur multiplicatif du temps réel. Les valeurs reportées sont des moyennes estimées à partir de l'encodage des signaux de parole en utilisant l'outil *profiler* de MatLab. Cette simulation a été réalisée en utilisant [MatLab®, 2011] sur une machine Linux utilisant un seul coeur cadencé à 2.5 Ghz. La durée totale des signaux utilisée est de 2 mins. Ce tableau donne une estimation de la complexité d'implémentation des différents blocs du codeur proposé. Par example l'algorithme de filtrage dans le bloc d'inversion des motifs peut traiter un signal de durée de 12.5 secondes en une seconde. Le tableau 6.3 montre que si on fait abstraction des blocs de compressions et

Tableau 6.3 Complexité computationnelle de l'implémentation proposée.

Bloc	Algorithme	Vitesse(temps réel)
Extraction des motifs	Filtrage	11.11
	Détection des pics	6.66
Masquage	Masquage pré-stimuli	0.12
	Masquage post-simuli	0.13
	Compensation d'énergie	0.12
Compression	RLE	0.53
	BWT	0.59
	Quantification	1.81
Inversion des motifs	Filtrage	12.50
	Égalisation	100.00

masquage le système proposé peut facilement opérer en temps-presque-réel[6] (*nearly real-time*) (vitesse valant 3 fois le temps réel). Ceci valide la simplicité computationnelle de l'approche proposée et l'idée originale derrière la conception des filtres binomiaux. L'implémentation peut être optimisée par example en combinant les blocs du masquage au bloc de la correction adaptative des amplitudes puisque ils se partagent les mêmes données. L'implémentation actuelle des ces algorithmes de masquage ne prend pas en compte la parcimonie des motifs d'excitation (les seuils de masquage sont estimés même pour les impulsions nulles dans le but de générer les figures du chapitre précédant). Sachant que

5. L'implémentation en C du même algorithme est 40 fois plus rapide.
6. Le délai algorithmique du système est de 25 ms.

ses motifs d'excitation sont à 82% épars, on s'attend à une réduction de complexité de même ordre de grandeur si les impulsions nulles sont exclues lors l'estimation du seuil de masquage. On peut aussi penser à réécrire ces algorithmes en un language de programmation bas-niveau par example en C. L'encodage des signaux de parole a pris en total 50 mins. Ceci reste rapide comparativement aux résultats reportés par [Thiemann, 2011] où il faut 30 mins pour encoder un signal dont la durée est de 10 secondes.

6.3 Conclusion

On a montré dans ce chapitre qu'il est possible de réduire le nombre de bits nécessaires à la transmission des motifs d'excitation auditives. Alors que la transformation BWT combinée au codage par plage permet de réduire par 80% le débit nécessaire pour la transmission des positions des impulsions masquantes, l'encodage des différences entre amplitudes permet de réduire par 70% le débit nécessaire à leur transmission. Combinant ces deux approches, le codeur proposé parvient à réaliser des taux de compression de l'ordre de 70% tout en gardant une bonne qualité subjective mesurée par l'algorithme PEAQ. Comparé aux autres approches opérant par codage des motifs d'excitations auditive, le codeur présenté dans cette thèse se distingue par sa complexité réduite tout en maintenant une bonne qualité de synthèse.

CHAPITRE 7

Conclusion générale

7.1 Contributions originales

Dans le chapitre 1 on a présenté les raisons qui ont motivé les travaux de recherche présentés par cette thèse : le paradoxe existant entre représentations auditives fidèles au fonctionnement de l'oreille humaine et la discipline du codage audio. Cette thèse est construite autour de ce dilemme où on démontre qu'il est possible à faible coût computationnel et à débit de transmission moyen de reconstruire un signal audio par inversion de ses motifs d'excitation auditive.

Banc de filtres auditifs à complexité réduite :

Le modèle le plus complexe en terme d'implémentation étant le banc de filtres d'analyse-synthèse, les premières parties de cette thèse s'attardent sur la conception d'un nouveau filtre auditif à faible complexité computationnelle. Le chapitre 2 introduit l'idée originale des filtres binomiaux où on détaille l'expression de leur formulation mathématique ainsi que les détails de leur implémentation digitale. On montre aussi dans le même chapitre, que ces filtres peuvent modéliser fidèlement les réponses impulsionnelles de la membrane basilaire de chats et on compare les résultats obtenus avec ceux publiés dans la littérature : les filtres binomiaux représentent un bon compromis entre fidélité de modélisation et coût d'implémentation digitale. Dans le chapitre 3, en se basant sur les expériences de masquage de [Baker *et coll.*, 1998; Glasberg et Moore, 2000], on démontre aussi que ces filtres peuvent modéliser le filtre auditif chez les humains. On compare aussi les résultats de simulations avec ceux des familles de filtres publiées dans la littérature et on montre que les filtres binomiaux peuvent être aussi utilisés pour modéliser le système auditif humain à moindre complexité. À titre d'example, l'implémentation digitale des filtres binomiaux nécessite 80% de coefficients en moins que les filtres Gammachirp.

Codage par inversion des motifs d'excitation auditive :

Vu que ce banc de filtres est simple à implémenter, dans le chapitre 4, on y juxtapose des modèles mimant le fonctionnement des cils ciliés et les neurones qui sont attachés pour extraire des motifs d'excitation auditive. On propose ensuite de synthétiser le signal original par inversion de ses motifs. On démontre dans ce chapitre qu'il n'est pas nécessaire

d'utiliser un banc de filtres de synthèse dont les réponses impulsionnelles sont celles du filtre d'analyse mais inversées dans le temps. En effet, on montre qu'il suffit d'introduire des lignes de retard pour être capable de synthétiser le signal original parfaitement. On donne aussi une méthode simple pour calculer la valeur de ses retards ainsi que les gains qui y sont associés. La réduction de la complexité introduite par toutes ces approches a permis de réaliser un codeur opérant à temps réel avec un délai algorithmique de 25 ms.

Masquage dans le domaine perceptuel :

Les motifs d'excitation auditive ainsi extraits sont épars mais restent quand même redondants : il y a en moyenne 5 impulsions par échantillon de signal de parole. Pour réduire cette quantité, dans le chapitre 5 on introduit des algorithmes de masquage opérant directement sur ces motifs. Dans le but de fournir un modèle à complexité réduite, les algorithmes de masquage proposés sont récursifs et donc l'estimation des seuils de masquage ne nécessite ni le stockage de motifs d'excitation unitaire ni une recherche exhaustive. Les résultats de simulations basés sur des métriques objectives et subjectives valident l'efficacité de ces algorithmes : on est capable de réduire considérablement le nombre d'impulsions à transmettre tout en maintenant une bonne qualité de synthèse. À 0.76 impulsions par échantillon, la différence objective de qualité moyenne du système proposé vaut -1.25 impliquant une bonne qualité de synthèse.

Compression des motifs d'excitation auditive :

Les experimentations réalisées dans le chapitre 6 montrent que le masquage dans le domaine perceptuel avantage la compression avec et sans perte des motifs d'excitation auditive. Étant donné que la récupération d'un signal à partir d'une représentation éparse quantifiée ne peut être achevée à bas débit [Goyal *et coll.*, 2008], on propose dans le chapitre 6 de compresser les positions des impulsions masquantes indépendamment de leurs valeurs. Pour ce faire, on combine la transformation de Burrows-Wheeler au codage par plage pour compresser les positions des impulsions masquantes. On démontre que c'est une approche efficace puisque cela permet de réaliser des taux de compression de l'ordre de 50% pour la transmission des positions des impulsions masquantes. Pour ce qui est de leurs valeurs, le codage des différences entre celles-ci donne les meilleurs résultats. Cela est confirmé par des simulations validées par des métriques subjectives. Combinant les deux approches on démontre qu'il faut aussi peu que 4.75 bits/échantillon pour reconstruire le signal original large-bande sans dégradations de qualité perceptibles. On finit le même chapitre par une discussion sur les détails d'implémentation et on démontre qu'il est en

effet possible de synthétiser un signal audio à partir de ses motifs d'excitation auditives compressés et ce à moindre coût computationnel.

7.2 Discussions et travaux futurs

Les motifs d'excitation auditive sont épars. On a proposé dans le chapitre 6 de compresser les positions des impulsions masquantes indépendamment de leurs valeurs. Goyal *et coll.* [2008] aborde la théorie de l'acquisition comprimée, *compressive sampling* (CS) de point de vue compression entropique et détaille les conditions une fois satisfaites garantissent la récupération du signal épars à partir d'un ensemble de mesures aléatoires. Il arrive à la conclusion décevante stipulant que si les positions ne sont pas transmises, il est impossible asymptotiquement de récupérer le signal épars. Il semble donc pour le moment, qu'on ne peut se passer de transmettre les positions des impulsions masquantes des motifs d'excitation auditive.

Une idée qu'on pourrait alors envisager dans ce cas, consiste à quantifier conjointement les positions des impulsions ainsi que la valeur de leurs amplitudes. Dans [Feldbauer, 2005] une approche basée sur la quantification conjointe de ces informations est proposée. Les auteurs proposent une approche combinant le codage par plage et la quantification vectorielle. Dans cette approche, les indices de la quantification sont introduits entre distances séparant les impulsions. De l'information secondaire reste pourtant nécessaire pour indiquer au décodeur le début de chaque segment. On pense que l'application conjointe de la transformation de Burrows-Wheeler aux indices quantifiés est une meilleure approche.

On peut aussi penser à exploiter la redondance entre différents filtres auditifs. C'est à dire utiliser des codeurs entropiques d'ordre plus élevés où on pourrait exploiter la notion de compression contextuelle [Lakhdhar et Lefebvre, 2012]. Ceci vient évidement avec un coût d'implémentation additionnel, mais on s'attend à réduire considérablement le débit de transmission. On finit par noter que la correction adaptative des amplitudes et leurs quantifications se font de façon indépendante. On pense qu'incorporer les deux blocs avantagerait la compression des motifs d'excitation auditive.

ANNEXE A

Annexe

A.1 Signaux de parole

Le tableau A.1 donne la liste des signaux de parole utilisés dans le chapitre 5.

Tableau A.1 Signaux audio TIMIT utilisés.

Entraînement				Évaluation			
Homme	Segment	Femme	Segment	Homme	Segment	Femme	Segment
DCD0	SI1415	EE0	SI2135	JBR0	SI1001	TG0	SI2162
GJ0	SI776	NKL0	SI1522	GK0	SI1952	EE0	SI2135
RLD0	SI964	JSK0	SI1052	PAR0	SI1576	PAB1	SI841
ESG0	SI702	PA0	SI1054	DCD0	SI1415	VH0	SI836
PAR0	SI1576	DAC1	SI844	VJH0	SI1556	PA0	SI1054
CAL0	SI1138	LAC0	SI2161	REB0	SI745	JSK0	SI1052
DS0	SI713	JSJ0	SI854	JD0	SI1937	BG0	SI1160
HG0	SX15	TG0	SI2162	DPK0	SI552	EDW0	SI1084
VJH0	SI1556	CDR1	SI556	KCL0	SX281	DRD1	SX284
JBR0	SI1001	GWR0	SI948	DS0	SI713	CDR1	SI556
DLC0	SI765	BJ0	SI815	DWH0	SI1925	NKL0	SI1522
JD0	SI1937	EEH0	SI1742	ESG0	SI702	JSJ0	SI854
GK0	SI1952	VH0	SI836	RLD0	SI964	EEH0	SI1742
DWH0	SI1925	ISB0	SI949	DLC0	SI765	ISB0	SI949
RLJ1	SI1671	ETB0	SI1148	HG0	SX15	BJ0	SI815
DPK0	SI552	PAB1	SI841	TR0	SI1933	LAC0	SI2161
KCL0	SX281	EDW0	SI1084	CAL0	SI1138	ETB0	SI1148
TR0	SI1933	DRD1	SX284	RLJ1	SI1671	DAC1	SI844
TN0	SI2324	RLL0	SI1514	GJ0	SI776	GWR0	SI948
REB0	SI745	BG0	SI1160	TN0	SI2324	RLL0	SI1514

L'ensemble contient 80 segments de signaux de parole subdivisés en deux classes équilibrées. Chaque classe contient vingt hommes et vingt femmes. Les interlocuteurs sont identifiés par leurs initiales suivies d'un chiffre pour les différentier quand il y a confusion due à des initiales communes. Les segments de paroles sont identifiés par un suffixe suivi d'un chiffre.

LISTE DES RÉFÉRENCES

3GPP (2007). *AMR speech Codec ; Transcoding Functions* (TS 26.090). 3rd Generation Partnership Project (3GPP).

Allen, J. (2001). Nonlinear cochlear signal processing. *Physiology of the Ear, Second Edition*, p. 393–442.

Allen, J. B. (2008). Nonlinear cochlear signal processing and masking in speech perception. Dans *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg, p. 27–60.

Andoh, M., Nakajima, C. et Wada, H. (2005). Phase of neural excitation relative to basilar membrane motion in the organ of corti : Theoretical considerations. *The Journal of the Acoustical Society of America*, volume 118, numéro 3, p. 1554–1565.

Aubury, M. et Luk, W. (1996). Binomial filters. *Journal of VLSI signal processing systems for signal, image and video technology*, volume 12, numéro 1, p. 35–50.

Baker, R. J., Rosen, S. et Darling, A. M. (1998). An efficient characterisation of human auditory filtering across level and frequency that is also physiologically reasonable. *Psychophysical and Physiological Advances in Hearing*, p. 81–88.

Baumgarte, F. (2001). A psychoacoustic model for audio coding based on a cochlear filter bank. Dans *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop*, IEEE. p. 139–142.

Bessette, B., Salami, R., Lefebvre, R., Jelinek, M., Rotola-Pukkila, J., Vainio, J., Mikkola, H. et Jarvinen, K. (2002). The adaptive multirate wideband speech codec (AMR-WB). *IEEE transactions on speech and audio processing*, volume 10, numéro 8, p. 620–636.

Békésy, G. V. (1953). Description of some mechanical properties of the organ of corti. *The Journal of the Acoustical Society of America*, volume 25, numéro 4, p. 770–785.

Boashash, B. (1992). Estimating and interpreting the instantaneous frequency of a signal. II : Algorithms and applications. *Proceedings of the IEEE*, volume 80, numéro 4, p. 540–568.

Brandenburg, K. (1999). MP3 and AAC explained. Dans *Audio Engineering Society Conference : 17th International Conference : High-Quality Audio Coding*. p. 1–12.

Brandenburg, K. et Bosi, M. (1997). Overview of MPEG audio : Current and future standards for low bit-rate audio coding. *Journal of the Audio Engineering Society*, volume 45, numéro 1/2, p. 4–21.

Britanak, V. et Rao, K. (2001). An efficient implementation of the forward and inverse MDCT in MPEG audio coding. *IEEE Signal Processing Letters*, volume 8, numéro 2, p. 48–51.

Burrows, M. et J. Wheeler, D. (1994). A block-sorting lossless data compression algorithm. *Digital Systems Research Center Research Reports*, volume 1.

Carney, L. H., McDuffy, M. J. et Shekhter, I. (1999). Frequency glides in the impulse responses of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, volume 105, numéro 4, p. 2384–2391.

Combettes, P. L. (1993). The foundations of set theoretic estimation. *Proceedings of the IEEE*, volume 81, numéro 2, p. 182–208.

Cooper, F. S. (1980). Acoustics in human communication : Evolving ideas about the nature of speech. *The Journal of the Acoustical Society of America*, volume 68, numéro 1, p. 18–21.

Cooper, N. et Rhode, W. (1997). Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea. *Journal of neurophysiology*, volume 78, numéro 1, p. 261–270.

Cooper, N. et Yates, G. (1994). Nonlinear input-output functions derived from the responses of guinea-pig cochlear nerve fibres : Variations with characteristic frequency. *Hearing Research*, volume 78, numéro 2, p. 221–234.

Crowley, J. L., Riff, O. et Piater, J. H. (2002). Fast computation of characteristic scale using a half octave pyramid. Dans *International Conference on Scale-Space Theories in Computer Vision*.

Dallos, P. (1996). Overview : cochlear neurobiology. Dans *The cochlea*. Springer, p. 1–43.

Dau, T., Puschel, D. et Kohlrausch, A. (1996). A quantitative model of the effectivesignal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, volume 99, numéro 6, p. 3615–3622.

Dayan, P. et Abbott, L. (2002). Neural encoding II : Reverse correlation and visual receptive fields. *Theoretical Neuroscience*, p. 62–80.

De Boer, E. et De Jongh, H. (1978). On cochlear encoding : Potentialities and limitations of the reverse-correlation technique. *The Journal of the Acoustical Society of America*, volume 63, numéro 1, p. 115–135.

de Boer, E. et de Jongh, H. R. (1978). On cochlear encoding : Potentialities and limitations of the reverse-correlation technique. *The Journal of the Acoustical Society of America*, volume 63, numéro 1, p. 115–135.

de Boer, E. et Nuttall, A. L. (1997). The mechanical waveform of the basilar membrane. I. frequency modulations (“glides”) in impulse responses and cross-correlation functions. *The Journal of the Acoustical Society of America*, volume 101, numéro 6, p. 3583–3592.

Decorsière, R., Søndergaard, P. L., MacDonald, E. N. et Dau, T. (2015). Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, volume 23, numéro 1, p. 46–56.

Der, R., Kabal, P. et Chan, W.-Y. (2003). Towards a new perceptual coding paradigm for audio signals. Dans *International Conference on Acoustics, Speech and Signal Processing*, IEEE. Volume 5. p. 1–457.

Duifhuis, H. (2004). Comment on "An approximate transfer function for the dual-resonance nonlinear filter model of auditory frequency selectivity"[J. Acoust. Soc. Am. 114, 2112–2117]. *The Journal of the Acoustical Society of America*, volume 115, numéro 5, p. 1889–1890.

Ekstrand, P. (2002). Bandwidth extension of audio signals by spectral band replication. Dans *Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio*.

Erfani, Y. (2016). *Applications of perceptual sparse representation (Spikegram) for copyright protection of audio signals*. Thèse de doctorat, Université de Sherbrooke, 136 p.

Feldbauer, C. (2005). *Sparse pulsed auditory representations for speech and audio coding*. Thèse de doctorat, Graz University of Technology, 160 p.

Feldbauer, C. et Kubin, G. (2003). Critically sampled frequency-warped perfect reconstruction filter bank. Dans *Proc. European Conf. on Circuit Theory and Design*, volume 3.

Feldbauer, C. et Kubin, G. (2004). How sparse can we make the auditory representation of speech ? Dans *8th International Conference on Spoken Language Processing*. p. 1–4.

Foster, B. et Herley, C. (1995). Exact reconstruction from periodic nonuniform samples. Dans *International Conference on Acoustics, Speech, and Signal Processing.*, IEEE. Volume 2. p. 1452–1455.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G. et Pallett, D. S. (1993). Darpa timit acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, volume 93.

Glasberg, Brian R. ; Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc*, volume 50, numéro 5, p. 331–342.

Glasberg, B. R. et Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, volume 47, numéro 1, p. 103–138.

Glasberg, B. R. et Moore, B. C. (2000). Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *The Journal of the Acoustical Society of America*, volume 108, numéro 5, p. 2318–2328.

Goyal, V. K., Fletcher, A. K. et Rangan, S. (2008). Compressive sampling and lossy compression. *IEEE Signal Processing Magazine*, volume 25, numéro 2, p. 48–56.

Greenberg, S. (1988). Acoustic transduction in the auditory periphery. *Journal of Phonetics*, volume 16, numéro 1, p. 3–17.

Griffin, D. W. et Lim, J. S. (1984). Signal estimation from modified short-time Fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, volume 32, numéro 2, p. 236–243.

Haddad, R. A. et Akansu, A. N. (1991). A class of fast gaussian binomial filters for speech and image processing. *IEEE Transactions on Signal Processing*, volume 39, numéro 3, p. 723–727.

Hartmann, W. M., Best, V., Leung, J. et Carlile, S. (2010). Phase effects on the perceived elevation of complex tones. *The Journal of the Acoustical Society of America*, volume 127, numéro 5, p. 3060–3072.

Healey, M. (1967). *Tables of Laplace, Heaviside, Fourier, and Z Transforms*. W. & R. Chambers.

Heinz, M. G. et Swaminathan, J. (2009). Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *Journal of the Association for Research in Otolaryngology*, volume 10, numéro 3, p. 407–423.

Hicks, M. L. et Bacon, S. P. (1999). Psychophysical measures of auditory nonlinearities as a function of frequency in individuals with normal hearing. *The Journal of the Acoustical Society of America*, volume 105, numéro 1, p. 326–338.

Holters, M. et Zölzer, U. (2015). GSTPEAQ–an open source implementation of the PEAQ algorithm. *Proc. of the 18th Int. Conference on Digital Audio Effects*.

Hukin, R. et Damper, R. I. (1989). Testing an auditory model by resynthesis. Dans *European Conference on Speech Communication and Technology*. p. 1243–1246.

Irino, T. et Kawahara, H. (1993). Signal reconstruction from modified auditory wavelet transform. *IEEE Transactions on Signal Processing*, volume 41, numéro 12, p. 3549–3554.

Irino, T. et Patterson, R. (2006a). A dynamic compressive gammachirp auditory filterbank. *Transactions on Audio, Speech, and Language Processing, IEEE*, volume 14, numéro 6, p. 2222–2232.

Irino, T. et Patterson, R. (2006b). Dynamic, compressive gammachirp auditory filterbank for perceptual signal processing. Dans *International Conference on Acoustics, Speech and Signal Processing.*, volume 5. p. 14–19.

Irino, T. et Patterson, R. D. (1996). Temporal asymmetry in the auditory system. *The Journal of the Acoustical Society of America*, volume 99, numéro 4, p. 2316–2331.

Irino, T. et Patterson, R. D. (1997). A time-domain, level-dependent auditory filter : The gammachirp. *The Journal of the Acoustical Society of America*, volume 101, numéro 1, p. 412–419.

Irino, T. et Patterson, R. D. (2001). A compressive gammachirp auditory filter for both physiological and psychophysical data. *The Journal of the Acoustical Society of America*, volume 109, numéro 5, p. 2008–2022.

Irino, T. et Unoki, M. (1998). A time-varying, analysis/synthesis auditory filterbank using the gammachirp. Dans *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE. Volume 6. p. 3653–3656.

ITU-BS-1116 (1997). Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. *International Telecommunication Union, Geneva.*

ITU-BS-1387 (2001). Method for objective measurements of perceived audio quality.

ITU-BS-1387 (2015). Method for objective measurements of perceived audio quality.

ITU-R BS. 1284-1, R. (2002). General methods for the subjective assessment of sound quality. *International Telecommunications Union.*

Jax, P. et Vary, P. (2004). Feature selection for improved bandwidth extension of speech signals. Dans *International Conference on Acoustics, Speech and Signal Processing*, IEEE. Volume 1. p. 1–697.

Jayant, N. S. (1974). Digital coding of speech waveforms : ăPCM, DPCM, and DM quantizers. *Proceedings of the IEEE*, volume 62, numéro 5, p. 611–632.

Kabal, P. (2002). An examination and interpretation of ITU-R BS. 1387 : Perceptual evaluation of audio quality. *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University*, p. 1–89.

Karjalainen, M. (1987). Auditory models for speech processing. *Proc. Int. Congr. Phon. Sciences. Tallinn.*

Katsiamis, A., Drakakis, E. et Lyon, R. (2007). Practical gammatone-like filters for auditory processing. *EURASIP Journal on Audio, Speech, and Music Processing*, volume 2007, numéro 1, p. 063685.

Koike, T., Shinozaki, M., Murakami, S., Homma, K., Kobayashi, T. et Wada, H. (2005). Effects of individual differences in size and mobility of the middle ear on hearing. *International Journal Series C Mechanical Systems, Machine Elements and Manufacturing*, volume 48, numéro 4, p. 521–528.

Kollmeier, B., Brand, T. et Meyer, B. (2008). Perception of speech and sound. Dans *Springer Handbook of Speech Processing.* Springer Berlin Heidelberg, p. 61–82.

Kubin, G. et Kleijn, W. B. (1999a). Multiple-description coding (MDC) of speech with an invertible auditory model. Dans *Speech Coding Proceedings, 1999 IEEE Workshop*, IEEE. p. 81–83.

Kubin, G. et Kleijn, W. B. (1999b). On speech coding in a perceptual domain. Dans *International Conference on Acoustics, Speech, and Signal Processing.*, IEEE. Volume 1. p. 205–208.

Lakhdhar, K. (2009). *Encodage entropique des indices binaires d'un quantificateur algébrique encastré*. Mémoire de maitrise, Université de Sherbrooke, 90 p.

Lakhdhar, K. et Lefebvre, R. (2012). Context-based adaptive arithmetic encoding of EAVQ indices. *IEEE Transactions on Audio, Speech, and Language Processing*, volume 20, numéro 5, p. 1473–1481.

Lee, H. Y., Raphael, P. D., Park, J., Ellerbee, A. K., Applegate, B. E. et Oghalai, J. S. (2015). Noninvasive in vivo imaging reveals differences between tectorial membrane and basilar membrane traveling waves in the mouse cochlea. *Proceedings of the National Academy of Sciences*, volume 112, numéro 10, p. 3128–3133.

Li, G.-L., Cho, S. et Von Gersdorff, H. (2014). Phase-locking precision is enhanced by multiquantal release at an auditory hair cell ribbon synapse. *Neuron*, volume 83, numéro 6, p. 1404–1417.

Li, H. et Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, volume 25, numéro 14, p. 1754–1760.

Lin, L., Holmes, W. H. et Ambikairajah, E. (2001). Auditory filter bank inversion. Dans *IEEE International Symposium on Circuits and Systems*, IEEE. Volume 2. p. 537–540.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, volume 28, numéro 2, p. 129–137.

Lopez-Poveda, E. A. et Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, volume 110, numéro 6, p. 3107–3118.

Lutfi, R. A. et Patterson, R. D. (1984). On the growth of masking asymmetry with stimulus intensity. *The Journal of the Acoustical Society of America*, volume 76, numéro 3, p. 739–745.

Lutzky, M., Schuller, G., Gayer, M., Krämer, U. et Wabnik, S. (2004). A guideline to audio codec delay. Dans *Audio Engineering Society 116th convention, Berlin, Germany.* p. 8–11.

Lyon, R. (2011a). Cascades of two-pole-two-zero asymmetric resonators are good models of peripheral auditory function. *The Journal of the Acoustical Society of America*, volume 130, numéro 6, p. 3893.

Lyon, R. F. (1983). A computational model of binaural localization and separation. Dans *International Conference on Acoustics, Speech and Signal Processing*, IEEE. Volume 8. p. 1148–1151.

Lyon, R. F. (1996). The all-pole gammatone filter and auditory models. *Acustica*, volume 82, p. 90.

Lyon, R. F. (2011b). A pole–zero filter cascade provides good fits to human masking data and to basilar membrane and neural data. *American Institute of Physics Conference Proceedings*, volume 1403, numéro 1, p. 224–230.

Maass, W. et Bishop, C. M. (2001). *Pulsed neural networks*. MIT press.

Makur, A. et Mitra, S. K. (2001). Warped discrete-Fourier transform : Theory and applications. *IEEE Transactions on Circuits and Systems I : Fundamental Theory and Applications*, volume 48, numéro 9, p. 1086–1093.

MatLab® (2011). *version 7.13.0.564 (R2011b)*. The MathWorks Inc., Natick, Massachusetts.

Matsui, T., Nakajima, C., Yamamoto, Y., Anho, M., Iida, K., Murakoshi, M., Kumano, S. et Wada, H. (2006). Analysis of the dynamic behavior of the inner hair cell stereocilia by the finite element method. *International Journal Series C Mechanical Systems, Machine Elements and Manufacturing*, volume 49, numéro 3, p. 828–836.

Meddis, R. et Lopez-Poveda, E. A. (2010). Auditory periphery : from pinna to auditory nerve. Dans *Computational models of the auditory system*. Springer, p. 7–38.

Meddis, R. et O'Mard, L. P. (2005). A computer model of the auditory-nerve response to forward-masking stimuli. *The Journal of the Acoustical Society of America*, volume 117, numéro 6, p. 3787–3798.

Meddis, R., O'Mard, L. P. et Lopez-Poveda, E. A. (2001). A computational algorithm for computing nonlinear auditory frequency selectivity. *The Journal of the Acoustical Society of America*, volume 109, numéro 6, p. 2852–2861.

Miao, L., Liu, Z., Hu, C., Eksler, V., Ragot, S., Lamblin, C., Kovesi, B., Sung, J., Fukui, M., Sasaki, S. *et coll.* (2011). G.711.1 annex D and G.722 Annex B-New ITU-T superwideband codecs. Dans *International Conference on Acoustics, Speech and Signal Processing*, IEEE. p. 5232–5235.

Millman, R. E., Johnson, S. R. et Prendergast, G. (2015). The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *Journal of cognitive neuroscience*, p. 533–545.

Moore, B. C. (1987). *Frequency Selectivity in Hearing*. Academic Press Inc.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C., Peters, R. W. et Glasberg, B. R. (1990). Auditory filter shapes at low center frequencies. *The Journal of the Acoustical Society of America*, volume 88, numéro 1, p. 132–140.

Moré, J. J. (1978). The Levenberg-Marquardt algorithm : implementation and theory. Dans *Numerical analysis*. Springer, p. 105–116.

Morris, O. J. (1995). MPEG-2 : where did it come from and what is it ? Dans *Proc. IEE Colloquium MPEG-2*. p. 1–5.

Ning, D. et Deriche, M. (2003). A bitstream scalable audio coder using a hybrid WLPC-wavelet representation. Dans *International Conference on Acoustics, Speech and Signal Processing*, IEEE. Volume 5. p. 1–417.

OPTICOM (2016). Opticom's PEAQ. `http://www.opticom.de/technology/peaq.php`.

Painter, T. et Spanias, A. (2000). Perceptual coding of digital audio. *Proceedings of the IEEE*, volume 88, numéro 4, p. 451–515.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, volume 59, numéro 3, p. 640–654.

Patterson, R. D. (1986). Auditory filters and excitation patterns as representations of frequency resolution. *Frequency Selectivity in Hearing*, p. 123–177.

Patterson, R. D. et Nimmo-Smith, I. (1980). Off-frequency listening and auditory-filter asymmetry. *The Journal of the Acoustical Society of America*, volume 67, numéro 1, p. 229–245.

Pépiot, E. (2015). Voice, speech and gender :. male-female acoustic differences and cross-language variation in english and french speakers. *Corela. Cognition, représentation, langage*, , numéro HS-16.

Pichevar, R., Najaf-Zadeh, H. et Mustiere, F. (2010). Neural-based approach to perceptual sparse coding of audio signals. Dans *Neural Networks, the 2010 International Joint Conference on*, IEEE. p. 1–8.

Pichevar, R., Najaf-Zadeh, H., Thibault, L. et Lahdili, H. (2011). Auditory-inspired sparse representation of audio signals. *Speech Communication*, volume 53, numéro 5, p. 643–657.

Pichevar, R., Rouat, J., Feldbauer, C. et Kubin, G. (2004). A bio-inspired sound source separation technique in combination with an enhanced FIR gammatone analysis/synthesis filterbank. Dans *Signal Processing Conference, 2004 12th European*, IEEE. p. 2063–2066.

Pratt, H. et Sohmer, H. (1976). Intensity and rate functions of cochlear and brainstem evoked responses to click stimuli in man. *European Archives of Oto-Rhino-Laryngology*, volume 212, p. 85–92.

Rosen, S. et Baker, R. J. (1994). Characterising auditory filter nonlinearity. *Hearing research*, volume 73, numéro 2, p. 231–243.

Ruggero, M. A. (1992). Physiology and coding of sound in the auditory nerve. Dans *The mammalian auditory pathway : Neurophysiology*. Springer, p. 34–93.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S. et Robles, L. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America*, volume 101, numéro 4, p. 2151–2163.

Salami, R., Laflamme, C., Adoul, J.-P., Kataoka, A., Hayashi, S., Moriya, T., Lamblin, C., Massaloux, D., Proust, S., Kroon, P. *et coll.* (1998). Design and description of CS-ACELP : A toll quality 8 kb/s speech coder. *IEEE transactions on Speech and Audio Processing*, volume 6, numéro 2, p. 116–130.

Schroeder, M. et Atal, B. (1985). Code-excited linear prediction (CELP) : High-quality speech at very low bit rates. Dans *International Conference on Acoustics, Speech and Signal Processing*, IEEE. Volume 10. p. 937–940.

Seneff, S. (1990). A joint synchrony/mean-rate model of auditory speech processing. Dans *Readings in speech recognition*, Morgan Kaufmann Publishers Inc. p. 101–111.

Shamma, S. et Lorenzi, C. (2013). On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *The Journal of the Acoustical Society of America*, volume 133, numéro 5, p. 2818–2833.

Slaney, M. (1995). Pattern playback from 1950 to 1995. Dans *Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century.*, IEEE. Volume 4. p. 3519–3524.

Slaney, M., Naar, D. et Lyon, R. (1994). Auditory model inversion for sound separation. Dans *International Conference on Acoustics, Speech, and Signal Processing.*, IEEE. Volume 2. p. II–77.

Tan, Q. et Carney, L. (1999). A phenomenological model for auditory nerve responses : including the frequency glide in the impulse response. Dans *Bioengineering Conference, 1999. Proceedings of the IEEE 25th Annual Northeast.* p. 23–24.

Tan, Q. et Carney, L. H. (2003). A phenomenological model for the responses of auditory-nerve fibers. II : Nonlinear tuning with a frequency glide. *The Journal of the Acoustical Society of America*, volume 114, numéro 4, p. 2007–2020.

Teachmeanatomy (2016). Teachmeanatomy : inner ear. `http://teachmeanatomy.info`.

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E. et Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network : Computation in Neural Systems*, volume 12, numéro 3, p. 289–316.

Thiemann, J. (2011). *A sparse auditory envelope representation with iterative reconstruction for audio coding.* Thèse de doctorat, McGill University, 167 p.

Thiemann, J. et Kabal, P. (2007). Reconstructing audio signals from modified non-coherent hilbert envelopes. Dans *Proceedings Interspeech.* p. 534–537.

Tohyama, M. et Koike, T. (1998). *Fundamentals of acoustic signal processing.* Elsevier.

Valin, J.-M. et Lefebvre, R. (2000). Bandwidth extension of narrowband speech for low bit-rate wideband coding. Dans *Speech Coding, IEEE Workshop 2000.*, IEEE. p. 130–132.

Verhulst, S., Dau, T. et Shera, C. A. (2012). Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission. *The Journal of the Acoustical Society of America*, volume 132, numéro 6, p. 3842–3848.

Wada, H., Miyamoto, D. et Sugawara, M. (2002). Active force generated by the motility of the outer hair cell. *International Journal Series C Mechanical Systems, Machine Elements and Manufacturing*, volume 45, numéro 4, p. 862–869.

Webster, J. C., Miller, P., Thompson, P. et Davenport, E. (1952). The masking and pitch shifts of pure tones near abrupt changes in a thermal noise spectrum. *The Journal of the Acoustical Society of America*, volume 24, numéro 2, p. 147–152.

Wolters, M., Kjorling, K., Homm, D. et Purnhagen, H. (2003). A closer look into MPEG-4 high efficiency AAC. Dans *Audio Engineering Society Convention 115*, Audio Engineering Society. p. 1–16.

Yeara Kozlov, T. W. (2013). Persistence1d. `https://github.com/yeara/Persistence1D`.

Zilany, M. S. et Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America*, volume 120, numéro 3, p. 1446–1466.

Zilany, M. S., Bruce, I. C., Nelson, P. C. et Carney, L. H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve : long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America*, volume 126, numéro 5, p. 2390–2412.

Ziv, J. et Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, volume 24, numéro 5, p. 530–536.

Zwicker, E., Feldlkeller, R. et Bosquet, J. (1982). Psychoacoustique : L'oreille, récepteur d'information. *Annals of Telecommunications*, volume 37, numéro 1, p. 110–132.

Zwicker, E. et Terhardt, E. (1974). Facts and models in hearing. volume 2, numéro 1, p. 26–30.